NAIST-IS-DT9561007

博士論文

オントロジーを利用した情報の共有化に関する研究

岩爪道昭

1998年7月27日

奈良先端科学技術大学院大学
情報科学研究科 情報処理学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
博士 (工学) 授与の要件として提出した博士論文である。

岩爪道昭

審査委員：　西田 豊明 教授
植村 俊亮 教授
松本 裕治 教授
武田 英明 助教授

オントロジーを利用した情報の共有化に関する研究*

岩爪道昭

内容梗概

インターネットの利用拡大によって、容易に入手可能な電子化情報は大規模・多様化してきている。それらの中には、情報源の目的が自明でなく、定義や構造が不明確な情報が数多く存在する。そのため、情報利用者、情報源の目的を理解すれば情報が得られるというわけではなく、利用者自らが能動的に情報を統合化するという立場が必要になる。単にデータを電子的に共有し、参照可能にするという意味ではなく、内容レベル、知識レベルでの情報共有を実現するためには、従来の情報検索技術のようなシンタックス的方法だけでは不十分であり、内容指向のアプローチが必要となる。

本研究では、オントロジーを用いた知的な情報共有を実現するための方法論を提案する。オントロジーは、「対象領域に関する概念と概念間の関係の記述」と定義され、内容に関係し、かつ部分的に形式化されているため、これまで述べて来たような非均質かつ大量の情報源を内容レベル、知識レベルで共有するための鍵となる概念である。

本研究では、(1) オントロジーを利用したネットワークからの情報の収集・分類・統合化法 (2) オントロジーを利用した工学的知識の組織化・共有化法 (3) オントロジーの構築支援方法の 3 点について議論を行ない、本研究で提案した各手法の有効性を検証するために、プロトタイプシステムを構築し、ネットワーク上のデータを用いた評価実験と考察を行なった。

(1) については、WWW に代表されるような多様性 (形式面、内容面)、分散性、大規模性を扱う情報源群を対象とした、オントロジーによる情報収集・分類・抽

*奈良先端科学技術大学院大学 情報科学研究科 情報処理学専攻 博士論文, NAIST-IS-DT9561007, 1998 年 7 月 27 日.

出システム IICA(Intelligent Information Collector and Analyzer) を実現した。このシステムは、ユーザの要求に応じて、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。また、収集した情報群をオントロジーに結び付け、体系的に分類・整理する。さらに、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、統合化を行なう。IICA の有効性を検証するために、WWW における評価実験を行った。その結果、オントロジーを中心にシステムを構築することにより、情報の収集・分類・統合化を一貫して行なうことができ、ネットワーク上の広範で不均質な情報源の利用を可能にすることがわかった。

(2) については、より詳細な背景知識の必要な技術情報の共有化に焦点を当て、ベテラン技術者が持っている経験やノウハウを後輩に継承するための枠組として ICoB(Intelligent Corporate Base) の考え方を導入し、変圧器改修計画業務支援を事例とした現場技術共有システム OnTheSpot を実現した。また、その有効性を検証するために、現場技術者による評価を行なった。オントロジーを用いることで、対象に関して全般的な知識を付加することができ、ドキュメントを対象の性質や構造などによって構造化して提示することで、当該の業務遂行に必要な背景的な知識を学習したり、対象の変化や追加に即したドキュメントを修正したりする際に有効であることがわかった。

(3) については、実データからのオントロジー構築支援の可能性について焦点をあてた。具体的には、(a)WWW ページのリンク情報、共起関係、類似度による概念間の関係の獲得法、(b) 情報量による概念の獲得法について、WWW ページを用いた獲得実験を行ない、得られたオントロジーについて検証した。その結果、(a) については、WWW ページのリンク情報、共起関係、類似度の 3 つ情報を用いることで、それぞれ違った側面の関係を見い出すことが可能であることがわかった。(b) については、獲得したオントロジーによる WWW ページの分類精度が、手動で作成した場合と同等であることがわかり、オントロジー作成支援が可能であることがわかった。

これらの結果から、オントロジーに基づくアプローチが、定義や構造が不明確で大規模な情報を、内容レベルで統合・共有化するのに有効な手段であることがわかった。

キーワード

オントロジー、WWW、情報収集、インターネット、イントラネット、現場技術情報

Studies On Ontology-based Information Sharing*

Michiaki Iwazume

Abstract

We are surrounded by enormous information supplied by electrical ways. Such information was once well-defined and well-organized, but nowadays much more ill-defined and poorly-organized information is supplied. We need ways to organize such information in order to obtain information related to our purpose.

One way to accomplish it is syntactical approaches like database views and text search. Database views approach is effective only where each information source is well-organized. Not only there is much variety of information sources, but also current information sources are themselves diverse and messy. Text search method is powerful and robust, but it becomes difficult to get meaningful results because of growth of information sources.

Content-based approaches are, thus, needed for such information sources instead of syntactical approaches. Ontology is a key concept for integration of heterogeneous information, because it commits ontology and partially formalized. In this paper, we propose ontology-centric knowledge organization approaches for integration and sharing of enormous and heterogeneous information.

First, we propose an ontology-based information gathering, classification, and extraction approach. We implemented a system called IICA(Intelligent Information Collector and Analyzer) which helps people to acquire knowledge from

*Doctor's Thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DT9561007, July 27, 1998.

information resources on the wide-are network. We tested IICA for tasks on the WWW. The result of the experiments indicated that the our approach enable us to use heterogeneous information resources on the wide-area network. We propose a supporting method of ontology acquisition using information extraction on the WWW.

Second, we propose an ontology-centric framework called Intelligent Corporate Base(ICoB)which realize a technical information sharing by supporting comprehension of documents. We implemented an prototype system called OnTheSpot for technical information sharing on repair work of trans. Then, some domain experts and field engineers validated its support function

and evaluated the performance of our method. They evaluated our method was useful enough in practical use.

Third, we propose three different methods to acquire ontologies automatically by using some statistical data of WWW pages. The first one is based on similarity of concept vectors generated by WWW page classification. The second is based on co-occurrence of terms in the generated concept vectors. The third is based on hyperlinks between classified WWW pages.

Furthermore, we developed a supporting method of ontology acquisition, which can choose candidate terms using category for discriminating pages. To examine our approach, we reclassified data using feature vectors whose elements were acquired by the method. The result showed that this method was better than a conventional method in classification of data and effective in ontology acquisition.

The above results indicate that it is possible to develop a system which helps people to share heterogeneous information resources on the wide-area network by using ontology.

Keywords:

ontology, WWW, information gathering, Internet, intranet, field engineeing information

目 次

図目次

表目次

第１章

序論

1.1　はじめに

インターネットの利用拡大によって、容易に入手可能な電子化情報は急増している。また、これらの情報は単に大量であるだけでなく、その定義や構造化の程度でこれまでの電子化情報と大きく異なっている。

かつての電子化情報は一定の目的のために作成され、提供されていた。このため、提供される情報は明示的な定義が存在し、組織化されていた。情報利用者は、情報提供者の目的を理解した上で利用していたので、情報の利用とは情報のアクセスと検索という問題に帰結していた。すなわち、利用者のするべきことは情報源に直接アクセスするか、あるいはアクセスの仕方をまとめたスキーマのようなものを作成して利用する程度であった。

WWW に代表される簡便な情報公開手段によって、ネットワーク化以前から蓄積された文書も含め、個人／組織レベルで作成された電子化情報がインフラレベルで共有されるようになってきている。

しかし、それらの中には、定義や構造が不明確な情報が数多く存在する。ここで問題になるのは、単に情報の規模や多様性ではなく、情報源の目的が自明でない点にある。WWW では、これまでの情報化の流れに沿って行なってきた我々の業務や生活の情報化した側面を情報源として利用することを可能にした。このことは、必ずしも情報源がその目的をもっているかどうか自明でないということを意味している。ある場合には暗黙的に目的が存在するし、またあるときはそもそも目的が存在しない場合もある。このような場合、当然情報利用者はこれまでのように、情報源の目的を理解すれば情報が得られるというわけではなくなる。こ

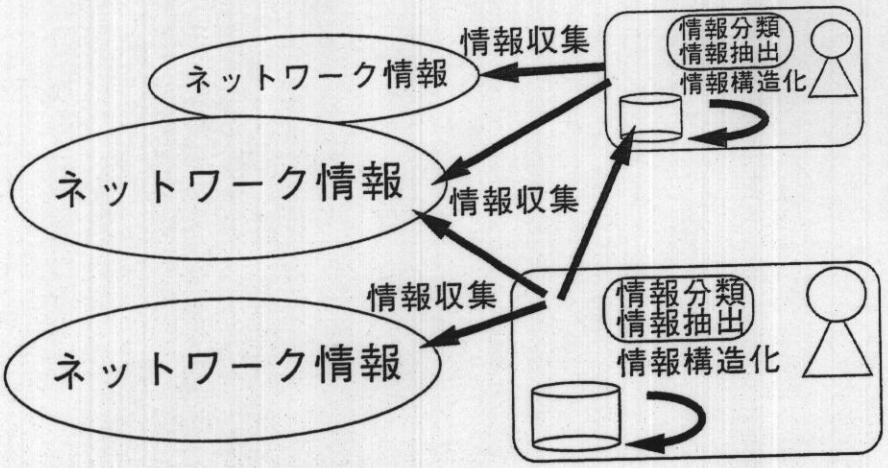


図 1.1 知的情報共有実現のための要素技術

のような受動的な立場ではなく、自らが情報を統合する作業をするという能動的立場が必要になる。そのためには、これまでのデータベースビューや情報検索といったシンタックス的方法だけではなく、内容指向の方法が必要になる。

本研究における情報共有とは単にデータを電子的に共有し参照可能にするという意味ではなく、内容レベル、知識レベルでのより知的な共有を意図している。

図 1.1は知的情報共有を実現するために必要な要素技術を摸式的に示したものである。知的情報共有は、情報収集と情報統合に大きくわけることができる。情報収集とはユーザの立場からどう情報源にアクセスするかという立場であり、従来の研究では情報検索、情報フィルタリングなどが関連する。

一方、情報統合としては、情報分類、情報抽出、情報組織化などが挙げられる。分類は収集してきた情報を適当なカテゴリーに分けることであり、クラスタリングなどの方法がこれまで行なわれてきた。情報抽出とは収集してきた情報から必要な情報を引き出すことであり自然言語理解などに関連する。また、情報構造化とは、さらにそれらの情報を関係づけを行なうことであり、発想支援などの研究分野と関わりがある。

本研究における知的情報共有では、これらの情報統合が情報収集と独立に行なわれるのではなく、お互いに影響しあって行なわれるところに特色がある。また、

情報統合を自分が行なうのではなく、他人が統合したものを利用する場合も考えられる。

1.2 オントロジーに基づくアプローチ

本研究では、前節で述べたネットワークにおける知的情報共有実現ための、オントロジー中心型情報共有化アプローチを提案する。広範囲な情報の関連性を知るためには単に情報源の情報を利用するだけでは難しく、背景的知識が必要になる。本研究では、背景的知識を提供する語彙体系、概念体系としてオントロジーの利用が不可欠であるという立場をとる。オントロジーは、内容に関係し、かつ部分的に形式化されているために、これまで述べて来たような非均質かつ大量の情報源を内容レベル、知識レベルで共有するのに鍵となる概念である。

また、情報収集、情報分類、情報抽出、情報組織化といった個々の技術をオントロジーを中心に構成することで、それぞれが独立に行なわれるのではなく、相互に影響しながら行なうことが可能になると考えられる。

1.2.1 オントロジーの定義

オントロジーは、元来哲学用語であり、「存在に関する体系的な理論」という意味で使われている。しかし、人工知能の分野では「概念化の明示的な仕様」[?]と定義され知識表現の規定として利用されることを意図している。ここで、概念化とは、対象世界に関する概念とそれらの間の関係を意味する。さらに、知識ベースの立場からは、「人工システムを構築するビルディングブロックとして用いられている概念／語彙の体系とその理論」という意味で使われている [?]。この定義は、知識の再利用を意識した定義といえる。

本論文ではオントロジーを、「対象領域に関する概念と概念間の関係の記述」と定義する。

1.2.2 オントロジーによる情報空間の表現

ここでは、前項で定義したオントロジーによって、対象の情報空間をどのように表現できるかという点について考察する。

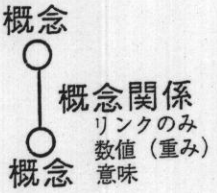


図 1.2 オントロジーの基本構造

オントロジーを構築する上で最も重要になるのは、対象領域に関する概念の切り出しとその意味の確定である。この段階で、必要な概念と関係が抽出され、それらの意味が定義され、オントロジーの本質が確定される。次に概念に関する合意が得られた後に、その概念に付与すべきラベルを決定する。

図 1.2は、前項のオントロジーの定義を最も単純にモデル化したものである。二つのノードは、それぞれ概念（語彙）を表す。概念は、通常、付与されたラベルで記述され、文字列で表現される。

リンクは概念関係を表している。概念関係には、(1) 単に関連性の有無だけを示すリンク構造を持つ場合、(2) 関連性の強さを示す数値をとる場合、(3) 上位・下位、IS-A などの意味を持つ場合があり、対象分野とオントロジーの利用法によって異なる。

本研究におけるオントロジーは、構造という観点では、グラフ構造、ネットワーク構造も取り得る。しかし、オントロジーの目的は、人間にも理解可能な形で情報を統合することであるため、ネットワーク構造を持つことは稀であり、2 章以降の研究でも階層構造で記述したオントロジーを利用している。

1.2.3 情報レベル

情報には、モデルレベル、形式表現レベル、メディアレベルの三つの階層があると考えられる。一番下の層は、モデルレベルであり、何らかのモデルで表現される情報を示す。ここでは、対象の記述方法が形式的に定義されていることと、対象の操作とその結果が定義される。これに対して中間の層は形式表現レベルであり対象の記述方法のみが形式的に定義される。いいかえれば、モデルレベルは

表現のシンタックスとセマンティックスが用意されているのに対して、表現レベルはシンタックスのみが定義されていることといえる。一番上のレベルはメディアレベルであり、形式的な定義もなく、情報の表現媒体の制約のみに規定される情報である。

1.2.4 情報共有におけるオントロジーの役割

情報共有におけるオントロジーの役割は、広範囲な情報の関連性を知るための背景的知識を提供することである。オントロジーが形式表現レベルのすべての機能を持つかどうかは議論の余地があるが、モデルとメディアをつなぐという機能は果たし得ると考えられる。オントロジーは計算機の中の記号システム、具体的には概念を表現する記号とその関係として記述される。人工知能でのオントロジーは、概念の同一性の維持など知識表現での合意を作ることを狙っていたので、オントロジーは概念世界の宣言的部分を主に表現している。このように、オントロジーは計算機の中の記号の体系というある種の形式性を持つが、セマンティクスは専ら人間の理解を通じて行なうだけでそれ自体は持たない。この意味で、モデルレベルとメディアレベル間の橋渡しするものとして適当であると考えられる。

以下では、本研究の中心課題である、情報収集・分類・抽出および情報構造化（組織化）という観点から、オントロジーの役割について考察する。

1．情報収集・分類・抽出におけるオントロジー　情報収集・分類・抽出におけるオントロジーの役割は、WWWページのように、同じ種類のメディアによって表現された莫大な情報群を、オントロジーの各概念に対応づけることによって統合することである（図 1.3参照）。

より具体的には、次のような役割があると考えられる。

1. 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、推論するための知識を提供する。

2. 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。

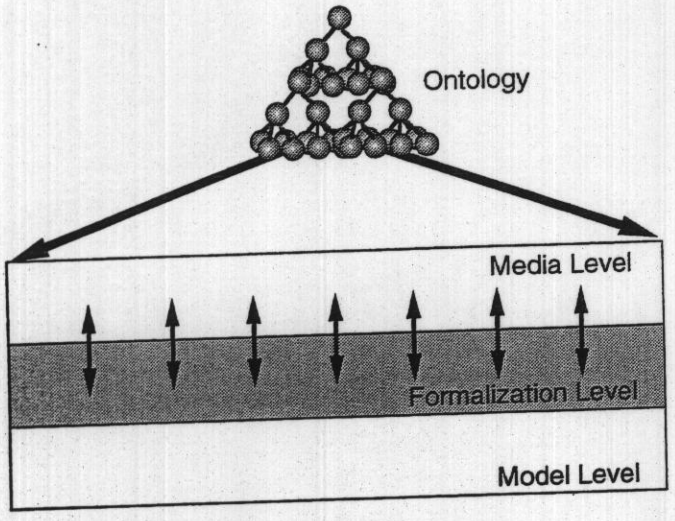


図 1.3 情報収集・整理のためのオントロジーの役割

3. 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。

4. 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を可能にする。

大規模な情報群には矛盾が多く含まれており、予め全てを考慮して体系的、網羅的に記述することは極めて困難である。そこで、情報収集・整理におけるオントロジー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に、クラス概念とその属性概念、概念間の関係のみで構成される。概念間の関係には様々な種類のものが考えられるが、ここでは、収集した情報の関連性を明らかにするのが目的であり、Ontolingua[?] のように論理的なものではない。

2. 情報構造化のためのオントロジー 情報構造化（組織化・体系化）では、情報収集の場合と比べて、より詳細な解が要求されため、異なるメディアやレベルの情報の関係を知る必要がある。

例えば工学的な問題解決である設計過程の場合、一般にただ一つモデルの利用で設計が完了することはなく、いくつものモデルを必要とする。その場合、モデ

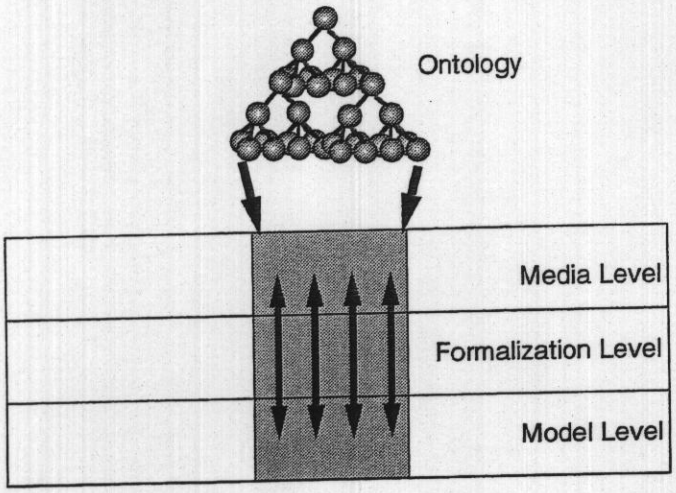


図 1.4 情報組織化におけるオントロジーの役割

ルとモデルの関係はより上位の形式表現レベルやメディアレベルを通じて行なっている。下位のレベルの情報の表現は、より領域に依存するが、その領域においては詳細に表現可能である。

したがって、ここでのオントロジーの役割は、図1.4に示すように、さまざまなメディアや異なるレベルの情報を統合することである。

1.2.5 オントロジーの獲得支援

従来、オントロジーの構築に関しては、オントロジー設計者が、構文と意味が厳密に定義された知識表現言語を用いて、できるだけ汎用で一般性のあるトピックを頭の中で考えながら記述する、という方法がとられてきた。

しかし、実データと独立して開発されたオントロジーは、設計者の内観によるトップダウン性が強く、ネットワークに散在する大規模かつ多様な情報源にそのまま適用可能である保証はない。また、厳密な定義を持ったオントロジーの構築は、多大な労力と時間を必要とする。

辞書の概念獲得やオントロジー構築や自動獲得に関する研究もすでに行なわれているが [?]、その多くは、汎用性を指向するあまり、どのように利用するかとい

う視点に欠けており、各研究で提案されている手法に基づいて獲得された辞書やオントロジーがどれほど有効なものであるか客観的に評価を行なっている例は少ない。

本研究では、オントロジーの利用を考慮に入れた、オントロジーの構築支援を考える。具体的には、オントロジーの獲得を (1) 概念の獲得と (2) 概念間関係の獲得に分け、WWW のリンク情報、共起関係、情報量等を用いたオントロジー獲得支援法について検討する。

1.3 研究の目的と方法

知的情報共有を実現するために、本研究では以下の 3 点を明らかにすることを具体的な目的とする。

1. オントロジーに基づくネットワークからの情報の収集・分類・統合化法を提案する
2. オントロジーに基づく工学的知識の組織化・共有化法を提案する
3. オントロジーの構築支援方法を明らかにする

また、以上の手法の有効性を検証するために、プロトタイプシステムを構築し、ネットワーク上のデータを用いた評価実験と考察を行なうことも、本研究の目的としている。

本研究では、以上の目的を実現するために、具体的に次のような方法を取る。

1. では WWW に代表されるような多様性 (形式面、内容面)、分散性、大規模性を扱う情報源群を対象とした、オントロジーによる情報収集・分類・抽出システム IICA(Intelligent Information Collector and Analyzer) を実現する。また、その有効性を検証するために、WWW における評価実験を行なう。

2. ではより背景知識の必要な技術情報の共有化に焦点を当てる。ベテラン技術者が持っている経験やノウハウを後輩に継承するための枠組として ICoB(Inteligent Corporate Base) の考え方を導入し、変圧器改修計画業務支援を事例とした現場

技術共有システム OnTheSpot を実現する。また、その有効性を検証するために、現場技術者による評価を行なう。

3. では 1 および 2 の研究で利用するオントロジーの自動構築がどこまで可能か議論する。具体的には、オントロジーの獲得を (1) 概念の獲得、(2) 概念間の関係の獲得に分け、WWW のリンク情報、共起関係、情報量等を用いたオントロジー獲得支援法について、WWW ページを用いた実験をもとに比較検討し、オントロジーの自動獲得がどこまで可能性かについて考察する。

1.4 本論文の構成

本論文は全 6 章から構成される。

第 2 章「オントロジーを利用した情報の収集・分類・統合化」では、ネットワーク上に散在する情報の共有を促進するために、オントロジーを用いた情報の収集・整理法について議論する。

第 3 章「オントロジーを利用した現場技術情報の共有化」では、工学的な問題における情報の組織化・共有に焦点を当てる。

第 4 章「オントロジーの獲得」では、第 2 章、第 3 章で利用したオントロジーの自動構築の可能性について議論する。

第 5 章「関連研究と考察」では、本研究で得られた成果の新規性、独自性、意義について関連研究との比較をしながら議論する。

第 6 章「結論」では、本研究の結論と課題について述べる。最後に情報共有のあり方について議論し、本研究の役割について述べる。

1.5 第 1 章のまとめ

本研究の目的は、オントロジーを用いた知的情報共有を実現するための方法論を明らかにすることである。情報源の情報を利用するだけでは、広範囲な情報の関連性を知ることは困難であり、背景的知識が必要になる。本研究では、背景的知識を提供する語彙体系、概念体系としてオントロジーの利用が不可欠であるという立場をとる。オントロジーは、「対象領域に関する概念と概念間の関係の記述」と定義され、知的情報共有の鍵となる概念である。本論文では、オントロジー

を利用したネットワークからの情報の収集・分類・統合化法、オントロジーを利用した工学的な知識の組織化・共有化法、およびオントロジーの構築支援方法を提案し、プロトタイプによる評価実験によって検証を行なう。

第 2 章

オントロジーを利用した情報収集・分類・統合化

2.1 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。そのため、我々がネットワークから必要な情報や知識を得るには、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。本章では、オントロジーを利用したネットワークからの情報の収集・分類・抽出・統合化法について議論する。

本章では, まず、インターネットに代表されるコンピュータネットワークが抱える課題の一つである情報の氾濫について述べる。また、現在さまざまなサイトでサービスされているサーチエンジンにおける特徴と問題点を挙げる。次に、オントロジーを用いた情報の収集・整理システム IICA (Intelligent Information Collector and Analyzer) を提案する。まず、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を示す。次に、オントロジーに基づく 情報の分類法について説明する。最後に、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。また、提案した各方法について、WWW を対象とした実験結果をもとに評価を行なう。

2.2 研究の背景

電子メディアや情報基盤の発達により、情報・知識は流通の過程で新たな情報・知識を生み出し、増大し続けている。例えば、インターネットでは、電子メール, メーリングリスト、ニュースグループ、Archie、Wais、Gopher、WWW 等のさまざまな情報サービスが、個人でも簡単に利用でき、世界中の情報にアクセスできるようになってきている。これは、情報技術者や研究者だけでなく、専門知識のない一般の人達も、情報生産活動に参加可能であることを意味している。

このような情報の多様化、大規模化は、個人の管理能力や従来の技術では対応しれない地点にまで達している。我々は、もはや利用可能なデータのうちごくわずかな部分しか利用できない情報過多の中に生きているといえる。

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なわれている。例えば、近年 WWW 上では、サーチエンジンと呼ばれる検索サービスが、急速にその数を増やしている。図 2.1は、日本の代表的なサーチエンジンについてまとめたものである。

2.2.1 サーチエンジンによる情報検索実験

サーチエンジンにおける問題点を明確にするため、本研究の対象領域の一つであるコンピュータ関連の製品情報に関する検索を図 2.1の九つのサーチエンジンを用いて、簡単な検索実験を行なった。検索キーワードは、

- “プリンタ” &” 価格”

である。ここで、” 価格” という言葉を検索キーワードとした理由は、” プリンタ” に関するページの中で、製品情報であるページには、価格情報があると判断したためである。表 2.1は、この二つのキーワードを入力した結果である。ここで、ヒット件数とは、入力キーワードに対する各サーチエンジンの検索結果のことである。Not Found は、実際にそのサイトにアクセスしたときに、そのページが存在しなかった件数を意味している。PRINTER のページ数とは、プリンタの製品情報について書かれており、正解と思われるページの件数のことである。ただし、検索方法で and 条件が指定できない場合は調べていない。

サーチエンジン	オープンテキスト	千里眼	ODIN	TITAN
アドレス	(http://www.jp.opentext.com/)	(http://www.info.waseda.ac.jp/search.html)	(http://kichijiro.c.u-tokyo.ac.jp/odin/)	(http://isserv.tas.ntt.jp/chisho/titan.html)
登録件数	非公開	非公開	22万件	非公開
ディレクトリによる検索	なし	なし	なし	なし
検索方法	AND,OR,BUTNOT,NEAR,HOLOWEDBY	AND	+,>,*(前方一致)	AND,OR
特徴	・検索結果の絞りこみ機能がある。 ・他のサーチエンジンに比べて、比較的新しい情報が検索できる。	・検索結果には、被リンク数やいつ存在確認したかの情報も表示去れる。 ・同一ディレクトリ内のページの除外も可能。	・検索結果を検索キーワードがヒットした語の数でスコアを求め、高い順に表示する。	・翻訳機能を持ち、海外のサイトも検索対象としている。 ・形態素解析 JUMANを使っている。 ・テキスト検索エンジンとしてWAISを使っている。
Yahoo!JAPAN	NETPLAZA	Hole-in-One	InfoseekJapan	mondou
(http://www.yahoo.co.jp/)	(http://netplaza.biglobe.or.jp/)	(http://hole-in-one.com/)	(http://japan.infoseek.com/)	(http://www.kuamp.kyoto-u.ac.jp/labs/infocom/mondou/index.html)
非公開	2万件	非公開	非公開	35万件
トップ・カテゴリ数14	トップ・カテゴリ数14	トップ・カテゴリ数16	トップ・カテゴリ数12	なし
AND,OR,*,+,#,?など使用可	AND,OR	AND,OR	OR,+,*(前方一致)	AND,OR,BUT
・ランダム・リンク機能がある。 ・ディレクトリ・サーチの場合の独自コメントがある。	・50音検索が可能出ある。 ・ NECの開発した、RetrievalExpressを使用している。 ・全文検索が可能である。	・文字検索の結果には、そのサイトの含まれるているカテゴリが表示される。 ・英文のページがあるサイトのみの表示が可能。	・海外の検索サービスの日本語版出。 ・検索キーワードの出現頻度、キーワード周辺のタグ情報を利用。	・検索結果の絞りこみ機能がある。 ・検索キーワードの関連語が表示され、絞りこみが可能。 ・合成語の自動分割機能画ある。 ・リンクの逆引が可能である。

図 2.1 日本のサーチエンジン

表 2.1 サーチエンジンの検索結果

サーチエンジン名	ヒット件数	Not Found	PRINTER のページ数
オープンテキスト	140	32	22
千里眼	10	1	1
TITAN	42	12	2
Yahoo! Japan	0	0	0
NETPLAZA	67	14	20
Hole-in-One	3	0	0
mondou	38	12	5

考察 ここでは、検索実験の結果をもとに、サーチエンジンにおける問題点を考察する。

実験に使用したサーチエンジンに共通の特徴・問題点があることがわかった。

(1) データベースが定期的に更新されている。

(2) ページの検索はキーワード入力によって行なわれる。

(3) 検索結果は、入力したキーワードの内容に最も近いと判断されたページから順にソートされて表示される。

(4) サーチエンジンによって得意な分野が存在する。サーチエンジンの特性を理解して利用する必要がある。

まず、(1) のデータベースの更新頻度については、コンピュータ関連の情報収集を行なう場合常に新しいデータが必要であるが、実験では、検索結果のリンクをたどっても実際にそのサイトにページが存在しない場合や更新されていない古いページが多数存在した。WWW サーバー数の拡大によって、サーチエンジンの更新スピードが遅くなってきていると考えられる。

ほとんどのサーチエンジンで、検索の問い合わせの方法として、(2) のようにキーワード入力を採用している。また、サーチエンジンによっては、AND、OR

だけでなく、BUTNOT(あるキーワードが存在しないページを対象とする) など論理的な検索ができるものもある。しかし、ユーザに検索の対象分野についての知識が不足していたり、論理的な検索方式に不慣れな場合、結果の絞り込みがうまくいかず、ユーザが望んでいる結果が得られない可能性が高いと思われる。

検索結果は、入力したキーワードの頻度等に基づく類似度計算によって (3) のようにソートされたページのリストが表示される。しかし、上位にランクされたページが、まったく関連のないページであったり、ヒット件数が膨大すぎて、全てのページ内容を調べることが困難である場合もある。

(4) については、表2.1からもわかるように、同じ問い合わせでも、ヒットする件数には違いがある。サーチエンジンによっては、ユーザが入力したキーワードを蓄積しそれに基づいてインデックスキーワードの追加を行なっているものがあり、データベースの規模の相違だけでなく、そのサーチエンジンを利用するユーザの興味に応じて、得意・不得意な分野が存在している。したがって、ユーザは自分の知りたい情報を検索するには、どのサーチエンジンを利用すればよいか、といったノウハウが必要になる。

2.2.2 ネットワーク上の情報氾濫による問題

サーチエンジンにおける実験でも明らかなように、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、情報の収集・理解・応用の段階で次のような問題が存在する。

1. **収集の問題:** 求める情報の所在が分からない。また, 専門知識の不足により求める情報そのものが具体的にわからない。

2. **理解の問題:** 収集してきた情報間の関係がわからない。また, 情報内容の背景知識が不足しているために, 理解するのに多大の時間と労力を要する。

3. **応用の問題:** 目的や要求に応じて, 保存した情報を取りだし, 有用な形に構造化することができない。

これらの問題を解決するには、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・整理できるようなより知的なシステムが必要である。

2.3 IICA の問題解決法

本章では、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・統合する IICA(Intelligent Information Collector andAnalyzer) と呼ぶシステムを提案する。

図 2.2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 2.3参照)。

2.3.1 IICA のオントロジー

オントロジーは, 概念化の仕様を記述したものである [?]。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, 全く何もない状態から, これらの言語を用いて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, 予めすべてを考慮して体系的, 網羅的に記述することは極めて困難である.

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オントロジーを採用する.

このオントロジーの構築には、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。

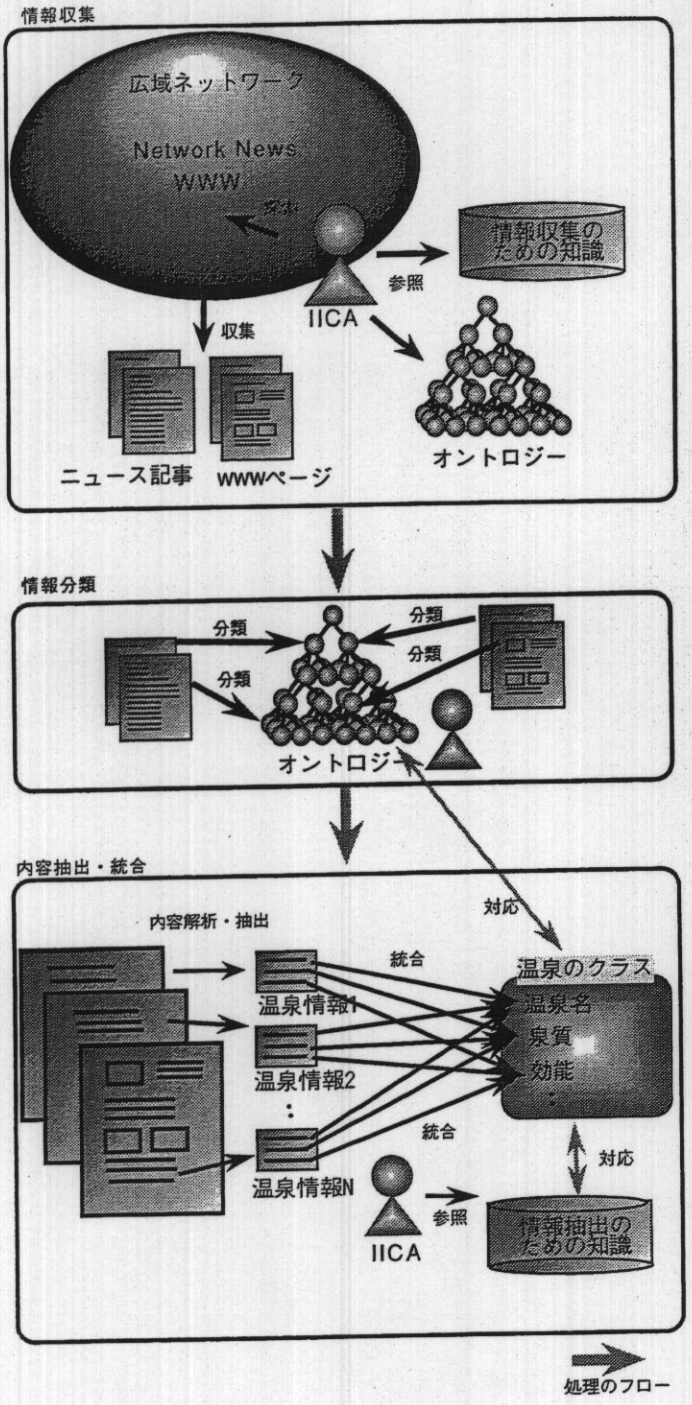

情報収集
広域ネットワーク
Network News
WWW
情報収集の
ための知識
参照
IICA
収集
ニュース記事
WWWページ
オントロジー
情報分類
分類
分類
分類
分類
オントロジー
内容抽出・統合
対応
内容解析・抽出
統合
温泉のクラス
温泉名
泉質
効能
温泉情報1
温泉情報2
温泉情報N
統合
対応
参照
IICA
情報抽出の
ための知識
処理のフロー

図 2.2 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"

図 2.3 温泉情報の統合例

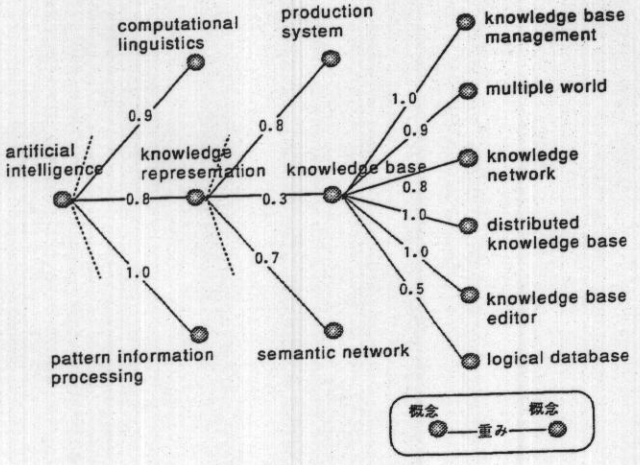


図 2.4 IICA のオントロジー

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 [?] 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、評価実験を行なっている。

また、図 2.4は、IICAで用いているAIに関する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リン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。オントロジーの利用法について次節以降で詳しく述べる。

2.4 オントロジーを用いた情報収集

オントロジーを利用したWWWにおける情報収集では、オントロジーによってアンカーをフィルタリングし、探索空間を絞り込むことが狙いである。以下では、収集アルゴリズムについて詳しく説明する。

2.4.1 収集アルゴリズム

探索は基本的に幅優先探索で行なう[1]。探索の実行には, 求める情報に関するキーワード、探索開始点となるURLアドレス、スコープパラメータ、収集ページ数の3種類の入力を必要とする。

システムは、入力したキーワードとの距離がスコープパラメータ以下にある概念をオントロジーからリストアップし、それぞれに対してキーワードの距離と同じ値の評価値（重み）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 2.5参照)。

[step1] 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、収集ページ数を入力する。

[step2] 入力したキーワードとの距離がスコープパラメータ以下の概念をオントロジーからリストアップする。

[1]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、といった対処が必要である。

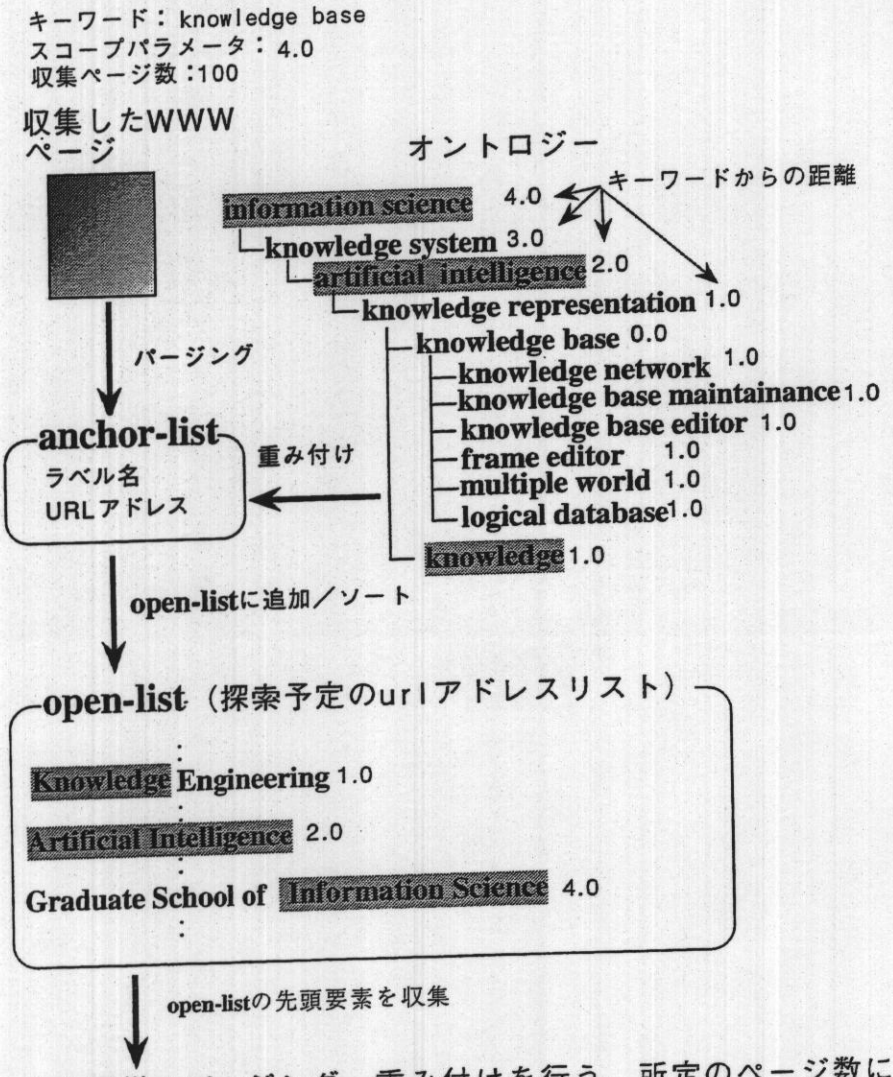


図 2.5 WWW における情報収集の例

[step3] 指定されたURLアドレスが既に収集したことがあるか調べる。新規のURLの場合は、HTTPにアクセスしてページを収集する。

[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了。満たなければstep5へ。

[step5] 収集したページのHTML構文を解析し、リンク部分に記述されているURLアドレス、見出しを抽出。各URLアドレスが既に収集済であるか、これから探索するURLアドレスのリストに含まれている場合は削除。それ以外は、step6へ。

[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どうか調べる。含まれているキーワードの重みからリンクに評価値を与える。これから探索するURLアドレスのリストに加え、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、これから探索するURLアドレスのリストから新しいアドレスを一つ取り出す。step3へ。

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2.5参照)。

IICAはまず、図2.5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。

ここで、説明を簡単にするためにオントロジーの各概念間の重みがすべて1.0であるとし概念間の距離を、

重み×概念間の最短経路のリンク数

と定義している。すなわち、親に相当する概念までの距離と子に相当する概念までの距離がすべて1.0となる。

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"knowledge" 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。

2.4.2 常識の利用

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。例えば、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。例えば、「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として与える。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ーを優先的に探索するように、重みを変更している。

2.4.3 評価実験

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なう。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二つの観点から行った。

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、収集実験を行なった。

a. 幅優先探索

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。知識（オントロジー、常識）は全く使用しない。

b. オントロジーの利用

幅優先探索の際、 2.3.1 項で説明した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

幅優先探索の際、2.3.1 項で説明したオントロジーおよび 2.3.2 項で説明した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 3 段階の基準に従い、著者が手作業で行なった。五つの問合せに対する評価の平均値を、表 2.2および表 2.3に示す。

ここでは、探索ステップ数を制限していないため、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大きな差が見られなかった。しかし、△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。このことから、ユーザが知りたい項目そのものではないが、それに関連する周辺のページを収集する際に、オントロジーが有効であることがわかった。常識については、ヒット率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。

表 2.2 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

表 2.3 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

表 2.4 収集効率の評価 — 一つのキーワード： “knowledge base”

探索法	○ (ページ数)	△ (ページ数)	× (ページ数)
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 2.5 収集効率の評価 — 二つのキーワード： (“semantic network” AND “production system”)

探索法	○ (ページ数)	△ (ページ数)	× (ページ数)
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

収集効率に関する検証　ここでは、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2.4は、一つのキーワード("knowledge base")からなる問い合わせに対して収集した結果、表2.5は、二つのキーワードの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、二つのキーワードによる実験(表2.5参照)では、1の方法で全く該当ページが収集できなかったのに対し、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を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。

2.5 オントロジーに基づく情報分類

本節では、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。

ここで、情報の分類とは、前節で説明した方法で収集したWWWページをカテゴリに分けることを意味する。すなわち、オントロジーに基づく情報分類では、オントロジーの各概念を分類のためのカテゴリとして用意し、収集したWWWページを各カテゴリに割り付けることである。 分類処理は大きく分けて次の過程からなる。

(1) WWWページの形態素解析／出現単語とその頻度からなるリストの生成

(2) WWWページの特徴ベクトルの生成／分類カテゴリの特徴ベクトルの生成

(3) WWWページの特徴ベクトルとカテゴリの特徴ベクトルとの類似度による、WWWページの分類先の決定

以下では、各プロセスについて詳しく説明する。次に分類精度向上のためのヒューリスティックスについて述べ、WWWページ分類評価実験を行ない、本アプローチの有効性について検討を行なう。

2.5.1 形態素解析と頻度リストの生成

各WWWページを形態素解析し、WWWページ中に出現する名詞とその頻度からなる頻度リストを生成する。WWWページはHTML形式で記述されているので、形態素解析を行なう前にHTMLのタグと文章を切り離す処理を行なっておく必要がある。WWWページの形態素解析には、日本語形態素解析システムJUMAN[?]を利用する。このシステムは日本語の文章を入力として、単語を切り分け、各単語の品詞情報を出力するものである。このシステムを用いて、WWWページ中から名詞を抽出する処理を行ない、さらにそれらの頻度を数えて、各ページに対する頻度リストを生成する。

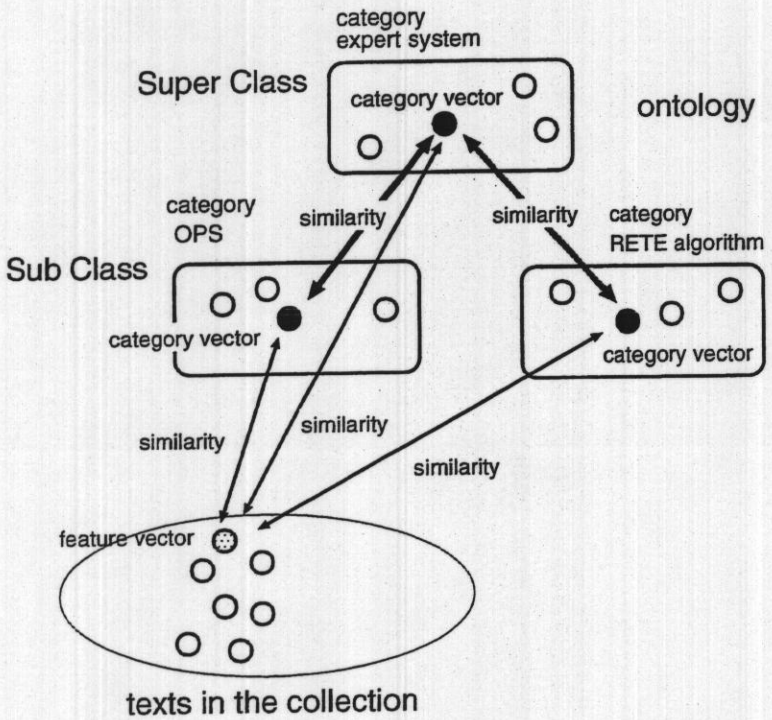


図 2.6 オントロジーによるテキスト分類

2.5.2 オントロジーに基づく特徴ベクトルの生成

特徴ベクトルは、対象領域に関する語の重みを要素として表現されたベクトルである。語の重みは、その出現頻度に応じて定義される。特徴ベクトルには、各WWWページに対するものと、オントロジーの各概念、すなわち分類のためのカテゴリに対するものの2種類がある。いずれのベクトルにおいても、その成分はオントロジー中の各概念に対応している。

一般に、オントロジーを構成する概念の集合 $(term_1, term_2, \ldots, term_t)$ に対して、あるオブジェクト $object_i$ の特徴ベクトル O_i は

$$O_i = (term_{i1}, term_{i2}, \ldots, term_{it})$$

である。ただし、$term_{ij}$ は $object_i$ における $term_j$ の重みを表す。

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 [?] 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。

WWW ページの特徴ベクトルの生成方法　WWW ページの特徴ベクトルは、基本的にはオントロジーの各概念に対応した成分が、それぞれの語の出現頻度を反映したものとなる。さらに、オントロジーを利用することによって、下位概念をもつ概念については、下位概念の出現頻度もベクトル成分に反映させる。具体的には、次のような操作を行なう。オントロジーの各概念をキーワードとし、先に作成したページの頻度リストとキーワードとのマッチングを行ない、マッチしたキーワードについて、その出現頻度を対応する成分の値とする。さらに、そのキーワードがオントロジーにおける上位概念を持つ場合は、上位概念に対応する成分にもその出現頻度を加える。こうしてできたベクトルを正規化したものを、その WWW ページの特徴ベクトルと定義する。すなわち、ある WWW ページ DOC_i の特徴ベクトル D_i は、

$$D_i = \frac{(term_{i1}, term_{i2}, \ldots, term_{it})}{\sqrt{\sum_{j=1}^{t}(term_{ij})^2}}$$

$$term_{ij} = (term_j\text{の出現頻度}) + (term_j\text{の子概念の出現頻度の総和})$$

で表される。

カテゴリの特徴ベクトルの生成方法　カテゴリはオントロジー中の概念に対応しているので、カテゴリの特徴ベクトルは概念の特徴ベクトルと言い換えることができる。カテゴリの特徴ベクトルは、対応する概念を表す単語が出現する WWW ページの特徴ベクトルを平均したものと定義する。すなわち、オントロジー中のある概念 $term_i$ のカテゴリに分類される WWW ページの集合を $(\mathrm{DOC}_1^i, \mathrm{DOC}_2^i, \ldots, \mathrm{DOC}_m^i)$ としたとき、$term_i$ に対応するカテゴリ CTG_i の特徴ベクトル C_i は、

$$C_i = \frac{\sum_{k=1}^{m} D^i{}_k}{m}$$

で表される。ただし、$D^i{}_k$ は DOC_k^i の特徴ベクトルである。

カテゴリの特徴ベクトルを用いた分類　カテゴリベクトルを用いた分類の手順を以下に示す。

[step1] 収集したページ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純なキーワードマッチングで分類する。

[step3] 分類されたページ群から各カテゴリの特徴ベクトルを計算。

[step4] 計算した各カテゴリ特徴のベクトルと収集ページの特徴ベクトルとの類似度を計算し、収集ページを再分類する。

[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。

2.5.3 類似度による分類先の決定

WWW ページとカテゴリの特徴ベクトルを用いて、WWW ページとカテゴリとの類似度を計算する。類似度は、WWW ページの特徴ベクトルとカテゴリの特徴ベクトルとの内積で定義する。すなわち、ある WWW ページ DOC_i と、あるカテゴリ CTG_j との類似度 $\mathrm{sim}(\mathrm{DOC}_i,\mathrm{CTG}_j)$ は、

$$\mathrm{sim}(\mathrm{DOC}_i, \mathrm{CTG}_j) = D_i \cdot C_j$$

で表される。各 WWW ページについて、すべてのカテゴリとの類似度を計算し、類似度が最大となるカテゴリに、その WWW ページを分類する。

2.5.4 精度を向上させるためのヒューリスティックス

前項で述べた方法に従い、旅行に関する WWW ページを対象として予備実験を行なった。あらかじめ収集した旅行に関する WWW ページ約 1000 件を本手法を用いて自動分類し、そのうちの無作為に選んだ約 400 件の内容を著者が実際に読んで、適切なカテゴリに分類されているかどうかを評価した。表?? はその評価結果を示している。実用的なシステムの構築を考えた場合、この分類精度はやや低い数字であると言える。分類の精度を向上させるためには、適切な分類をすることができなかったページの典型的な例を調べ出し、それらを適切に分類できるようにするための方法を考える必要がある。そこで、分類の精度を向上させる可能性があると考えられるいくつかのヒューリスティックスについて考察していくことにする。

表 2.6 予備実験の評価

正し分類	他の内容が混在	誤った分類
57.9%	25.8%	16.3%

インデックスページの除去　特徴ベクトルを用いた分類が困難なページの典型的な例として、インデックスの役割を果たしているページが挙げられる。インデックスページは、異なるカテゴリに属すると思われる種々のページへのリンクを持っており、それ自体は比較的特徴のないページである。したがって特定のカテゴリに分類することは難しい。インデックスページそのものは、他のページへのリンク以外の情報をほとんど持たないので、そこからリンクをたどって得ることのできるページを収集の段階で獲得できていれば、むしろインデックスページを分類の対象外とすることが望ましいと考えられる。

インデックスページの識別には、ハイパーリンクの数によるヒューリスティックによって、「案内型尺度」と呼ばれる離散値をページに付与する方法が考えられている [?]。このような方法を用いて分類を実行する前に予めインデックスページを除去することができれば、分類の精度の向上につながと考えられる。

複数のカテゴリへの分類　複数のカテゴリに属すると思われる内容が一つのページに含まれているものも、分類が困難な例の一つである。このようなページは、内容ごとに分割してそれぞれ該当するカテゴリに分類することが望ましいが、ページを分割するとその内容が不明確になってしまう場合もあり、またページのどの部分で内容が変わるかを機械的に判断することは困難である。

そこで、ページを分割する代わりに、一つのページを複数のカテゴリに分類することを許す、という方法が考えられる。類似度に適当な閾値を設け、これを超えるカテゴリについては無条件にそのページを割り当てる、というものである。この方法によって分類結果の冗長さが増すという欠点はあるが、欲しい情報がより多く得られる可能性があるという点では有用である。

対象領域に無関係なページの除去　本研究におけるWWWページの分類では、予め対象とする分野を定め、その分野に関するページだけを収集し、得られたページを分類の対象とする、ということを前提としている。しかし、ページの自動収集を考えた場合、収集したページが全て対象としている分野に関するものであるというわけにはいかず、実際には収集されたページの中に対象領域に無関係なページが存在する。

一般には、対象領域に無関係なページならば、対象領域のオントロジーが持つ語彙とマッチする語が全く存在しないことによってそのページを除外できる、と考えられる。しかし、オントロジー中に他の分野にも関連のある概念が存在する場合もあり、そのような概念とマッチすることにより、無関係なページが誤って分類されてしまう可能性がある。例えば、旅行に関するページを対象とした分類実験では、旅行関連以外でもよく使われる一般的なキーワード（例えば「文化」「歴史」など）とマッチして、旅行に関係ないページを分類してしまう場合があり、分類の精度を低下させる一因となっている。

2.5項で述べた方法では、キーワードにマッチする単語がページ中に一つでもあれば、必ずいずれかのカテゴリに分類されることになっている。しかし、対象領域に無関係なページであれば、その分野に関するキーワードが出現する割合は非常に低いと考えられる。したがって、ページ中の全単語数（名詞のみ）に対するキーワードの出現率を計算し、値が著しく低いものは旅行に無関係なページとして除去することができると考えられる。

カテゴリベクトル生成における例外的なページの除外　カテゴリの特徴ベクトルの初期値の生成は、そのカテゴリを表す用語が出現するページを集め、これらのページの特徴ベクトルの平均をとることによって行なっているが、その用語以外のキーワードが全くないページの場合、そのページの特徴ベクトルは唯一のキーワードに対応する成分が1で、他の成分はすべて0のベクトルとなる。このようなページはそれ自身がもつ情報量が非常に少なく、例外的なものであり、そのようなページのベクトルは他のページのベクトルとの相違が大きい可能性がある。したがって、このような例外的なベクトルはカテゴリのベクトルの計算の対象外とするのが妥当であると考えられる。この考えを取り入れることによって、カテ

ゴリベクトルがそのカテゴリに分類されるべきページの特徴をより忠実に表すものになると考えられる。

分類結果からのフィードバックによるキーワードの追加　その他のうまく分類されない例として、旅行に関するページであるにもかかわらず、マッチする単語が存在しないためにどのカテゴリにも分類されないページがある。このようなページのなかには、名所の名前（固有名詞）と写真だけのページというような、キーワードにマッチさせることが不可能なものもあるが、オントロジーの語彙が不足していることが原因で分類に失敗しているものも存在することがわかった。そこで、分類されたページの内容をみていったとき、必要と思われるキーワードを追加することによって、結果として分類の精度か向上することが期待される。

2.5.5　評価実験

AI（knowledge base）に関して収集した約500件のWWWページおよび、旅行全般に関して収集した約800件のWWWページに対して、分類実験を行なった。ただし、旅行に関するページについては、予め収集した約1000件のうち、インデックスページ、旅行に無関係なページを手作業で除去し、残りの約800件に対して実験を行なった。

分類精度の評価は、情報検索において広く用いられている、適合率（分類ノイズの少なさ）と再現率（分類もれの少なさ）を以下の式を用いて求めた。表??に結果を示す。

$$\text{適合率} = \frac{\text{正しく分類されたページ数}}{\text{カテゴリに分類されたページ数}} \cdot 100\ (\%)$$

$$\text{再現率} = \frac{\text{正しく分類されたページ数}}{\text{カテゴリに分類されるべきページ数}} \cdot 100\ (\%)$$

英文のAIに関するページ、和文の旅行に関するページで約80%程度の適合率が得られた。再現率に関しては、和文の旅行関連ページの方が70%と若干精度が落ちている。

この原因の一つとしては，日本語の形態素解析が不十分であることまずあげらる。また、その他の要因としては、旅行という分野の特殊性が考えらる。これは、

表 2.7 WWW ページ分類実験の評価

	AI のページ (英語)	旅行のページ (日本語)
適合率	81.9%	79.0%
再現率	80.5%	70.0%

AI, 情報科学全般などの学門分野で扱う概念は、包含関係が比較的明確であるのに対し、旅行では各概念の包含関係が不明確であったり、歴史や自然といった曖昧で特徴を見出しにくい概念なども取扱っているということである。

そこで、以下では旅行のページに関する分類結果の詳細についてさらに詳しく検討してみる。表??は、旅行データに対する分類実験で得られた主なカテゴリの適合率、再現率を示している。

「祭」「寺」「温泉」などのような、具体的な概念については、高い適合率が得られている。これに対し、「自然」「歴史」といった抽象概念は、やや低い適合率となっている。

「交通」に誤って分類されているページの多くは具体的な観光地についてのページであった。これは、観光地についてのページの場合、関連語まで含めると1ページ当たりの「交通」の出現頻度が非常に高くなり、特徴ベクトルにおける「交通」の成分の値が大きくなるためであることがわかった。確かに、観光地についてのページにおいて交通は重要な要素と考えられるが、このようなページは例えば「遊園地」、「美術館」といった具体的な行き先を表すカテゴリに分類されることが望ましく、「交通」の記述による影響を抑制するための手段が必要である。このための解決策としては、現在のところすべて同等に扱っているキーワードに対して、収集ページ全体における出現頻度が高いほど低い重みを与えるようにする方法が考えられる。

「山」に誤って分類されているページは、主として固有名詞の一部に「山」という文字が入っているものであった。すなわち、形態素解析の際にこの1文字が切り分けられて、単語と認識されている。形態素解析における固有名詞辞書が拡充されれば、この問題を解決することができるものと思われる。

表 2.8 主なカテゴリの適合率と再現率

カテゴリ	適合率 (%)	再現率 (%)
宿泊・ホテル・旅館	100	63
交通	47	66
見所	100	66
観光	89	74
行事	100	70
祭	100	80
自然	52	78
山	45	87
公園	84	81
文化	88	81
(美術 | 博物 | 科学 | 資料) 館	96	94
歴史	71	100
神社	96	59
寺	90	65
温泉	94	76
店・レストラン	100	33
食事・料理	94	76
施設	51	71

「店」の再現率が低いが、これは「施設」「料理」に誤って分類されるものが多数あることによるものであった。ただし、本実験で使用した「店」に関するWWWページのほとんどは飲食関係のものであり、「料理」に分類されることは必ずしも誤りではないと考えられる。また、「施設」は「店」の上位概念と考えることができるので、見方によっては「店」の再現率の低さはそれほど深刻なものではないと言える。

高い適合率を示している、「神社」「寺」「温泉」などの具体的な目的地を表すカテゴリの再現率が低くなっているが、これは本来これら目的地のカテゴリに分類されるべきページの多くが交通に分類されてしまうことによるものであった。

2.6 オントロジーとヒューリスティックスを用いた情報抽出

本節では、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たテキストからの情報抽出法について述べる。

本研究における情報抽出とは収集・分類したWWWページから、ユーザが必要な情報に関する記述部分を抜き出すことである。情報抽出は文書から情報を取り出すという意味では、自然言語処理であるが、本研究では、いわゆる文書理解を行なうのではなく、必要な情報だけを効率的に取り出す点に重点をおいている。ネットワーク上に情報のように多種多様な情報がある場合は、正確な理解よりもユーザが欲する情報があるかないか、あるとすれば何かといった粗い理解が必要である。

本研究では、ページの内容を自動抽出する手法として、以下の二つを提案する。

1. 状態遷移図を用いた方法　状態遷移図を用いて文章の内容を理解していく手法である。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　概念 (温泉、寺など) がテキストではどの

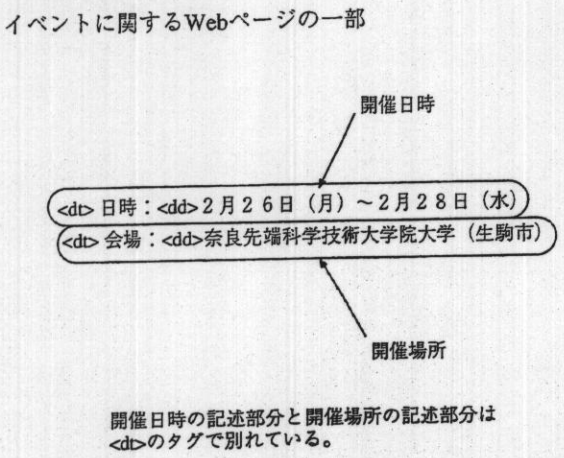


図 2.7 内容の差異によって文単位に分割

ように記述されているかをルールとして設定し、その設定したルールをテキストに適用していくという手法である。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、第??項ではこの両者の手法に共通する前処理について説明を行ない、第??項では状態遷移図を利用した手法を、第??項では概念の記述ルールを利用した手法について説明する。

2.6.1 HTML の基本処理

本研究では WWW ページを対象としているので、HTML フォーマットに対する処理を考える必要がある。そこで、IICA では 1) 文単位への分割、2) 単語単位への分割の二つの処理を行なうことにした。

文単位への分割 本研究では記述の内容が変わる箇所で分けている。つまり、分割された各々の記述部分が一つの内容のみを含むようにする。本研究では、分割された記述部分の一つひとつを文と呼ぶことにする。記述内容の差異によって

表 2.9 文単位への分割に関する基準

(1)「<Hn>～</Hn>」のパターンで記述してあれば、 それを一つの文と判断 (n = 1 or 2 or 3) する。
(2)「<DT>～</DL>」、「<DT>～<DT>」、「<DT>～<P>」、 「<DT>～<DL>」、「<DT>～<UL>」のパターンで記述して あれば、それを一つの文と判断する
(3)「<LI>～</UL>」、「<LI>～<LI>」、「<LI>～<P>」、 「<LI>～<DL>」、「<LI>～<UL>」のパターンで記述してあれば、 それを一つの文と判断する。
(4) その他のタグについては、文とみなさず無視する。
(5) その他の場合は、通常の文章のように「。」を文の区切りとする。

WWW ページを文単位に分割することで、後の情報抽出の処理を容易にするためである。次に、欲しい情報が記述されている文だけを抽出する。例えば、図??において、開催場所と開催時間に関する文に分けることで、開催場所の記述だけを抽出が可能になる。個々の文が一つの内容のみを含むように分割するのは困難である。本研究では表??を基準に WWW ページを文単位に分割している (具体例は図??参照)。

単語単位への分割 IICA では、文字列照合によって柔軟な情報抽出を行なうために、JUMAN[?] を用いて形態素解析をしている。しかし、形態素解析をすることによって文を単語単位に分けることはできるが、人間が単語単位に分けるのと同じ結果が得られるとは限らない。そこで、本研究では JUMAN によって形態素解析した後、表??のルールを用いて単語を連結させることで、人間が意図しているのに近い形態素解析ができるようにしている (具体例は図??参照)。

(1) の数字が隣接している場合は、それらが一つの数を表しているので、一つの単語と判断するようにしている。

(2) 数字の直後に単位が出現する場合にそれらを連結して一つの単語としたのは、例えば「数字の直後に km という単位がある」とルールを作っておけば、単語単位にルールとの照合ができるためである。「10km」が単語として存在してい

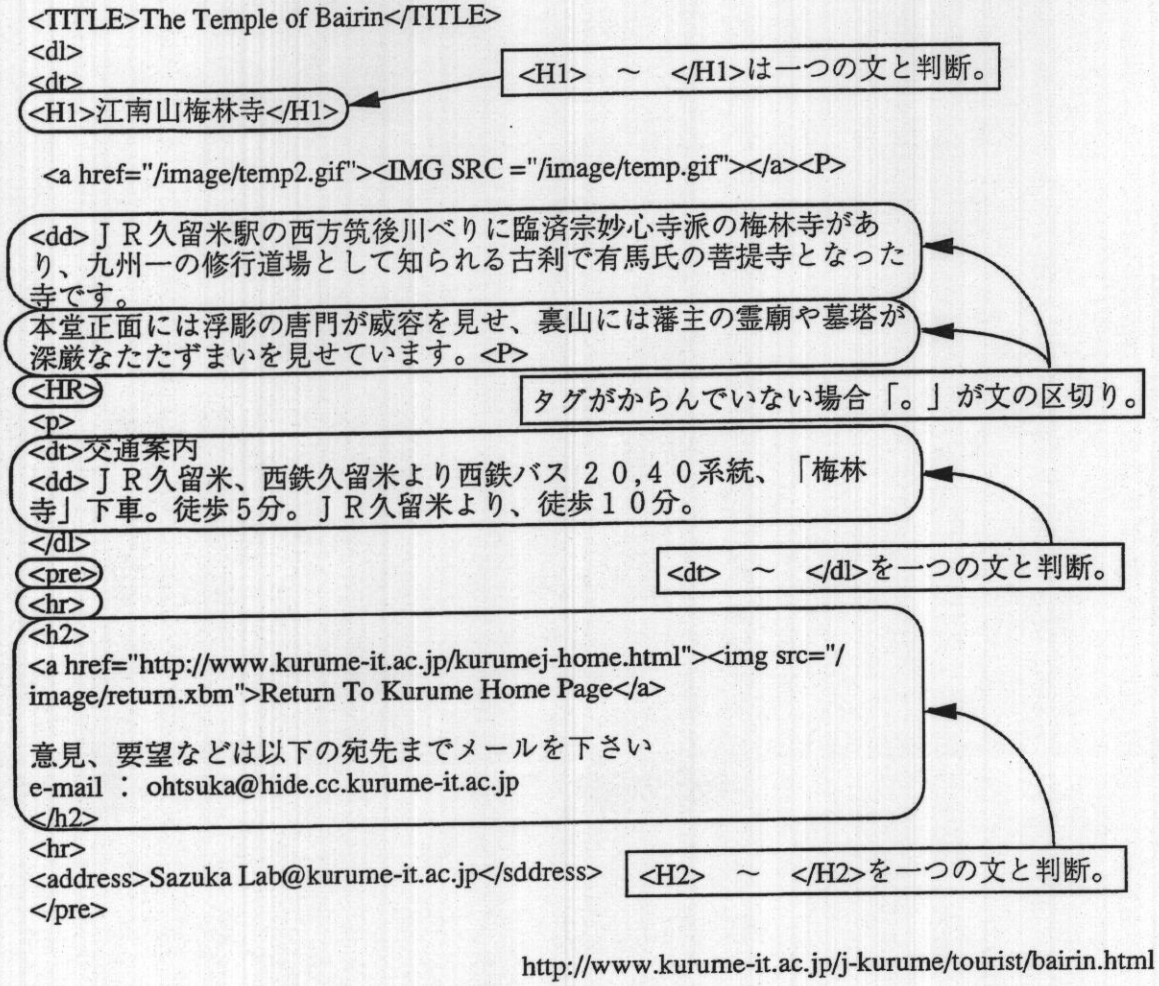


図 2.8 文単位の分割処理

表 2.10 形態素解析の後処理に関するルール

(1) 数字が隣接している場合、それらを連結する
(2) 数字の直後に単位が出現する場合、それらを連結する
(3) アルファベット同士が隣接している場合、それらを連結する
(4) 固有名詞または普通名詞が隣接している場合、それらを連結する

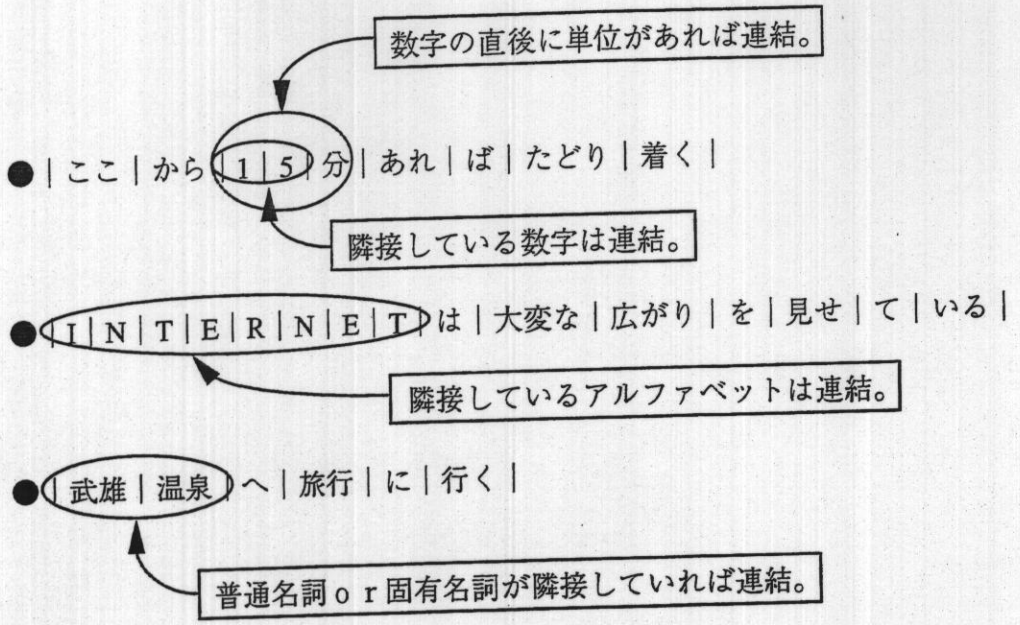


図 2.9 単語の切り分け

たとすれば、上述のルールに適合し、ユーザが必要とする情報に関する記述部分が抽出できる。

(3) のアルファベット同士が隣接している場合は、(1) と同様のことがいえ、それらが一つの単語を表している場合が多い。よって、アルファベット同士が隣接している場合はそれらを連結して一つの単語としている。

(4) の固有名詞もしくは普通名詞が隣接している場合は、それらを連結して一つの単語としたのは以下の理由からである。例えば「大阪駅」は形態素解析の結果、「大阪」という固有名詞と「駅」という普通名詞の二つに分解されるが、これでは「大阪駅」という情報が失われ、正確な情報抽出が行えない。そこで、このようになる単語は形態素解析の後に連結するよう処理している。

固有名詞もしくは普通名詞が隣接していれば、それらを連結して一つの単語としていると述べたが、連結して一つの単語としない方がいい場合もある。例えば「東海道本線京都駅」の場合、「東海道本線」と「京都駅」の二つに分けるのが自然であるが、上述の処理を行えば「東海道本線京都駅」と一つの単語になる。そ

表 2.11 形態素解析の後処理の特別なヒューリスティックス

(1)「～線」もしくは「～駅」の後には単語を連結させない。
(2)「下車」もしくは「徒歩」は他の単語と連結させない。

こで本研究では、連結して一つの単語にした時に不都合であることが分かる場合についてのみ、それらは連結させないでおくというヒューリスティックスを与えている（表??参照)。

2.6.2 状態遷移図を利用した手法

この手法は、「ページの中で、特定の単語の後にどのような単語が出現するか」が経験的に予測できる場合に用いる。これより具体例をあげて、この手法について述べる。

本手法の手順 この節では、ある観光地の WWW ページを例に本手法の手順を説明する。図??は交通手段記述の解釈を行なうための状態遷移図である。この状態遷移図を利用して観光地の交通手段情報を抽出することにする。(図??参照)。

[step1]WWW ページを文単位に分割し、文単位で情報抽出が行えるようにする (図??の (1))。

[step2] 文単位に分割した後、求める情報 (この例では交通手段) が記述してある文を抽出する (図??の (2))。

[step3] 抽出した文に対して形態素解析を行なう (図??の (3))。文を単語単位に分割にすることで、状態遷移図を利用した記述解釈ができるようになる。形態素解析は 3.1.1 項で述べた規則に従って行なう。ここまでは、状態遷移図を利用した手法の前処理である。

[step4] 対象とする文中のどの単語から状態遷移図による解釈処理を始めるかを決定する (図??の (4))。

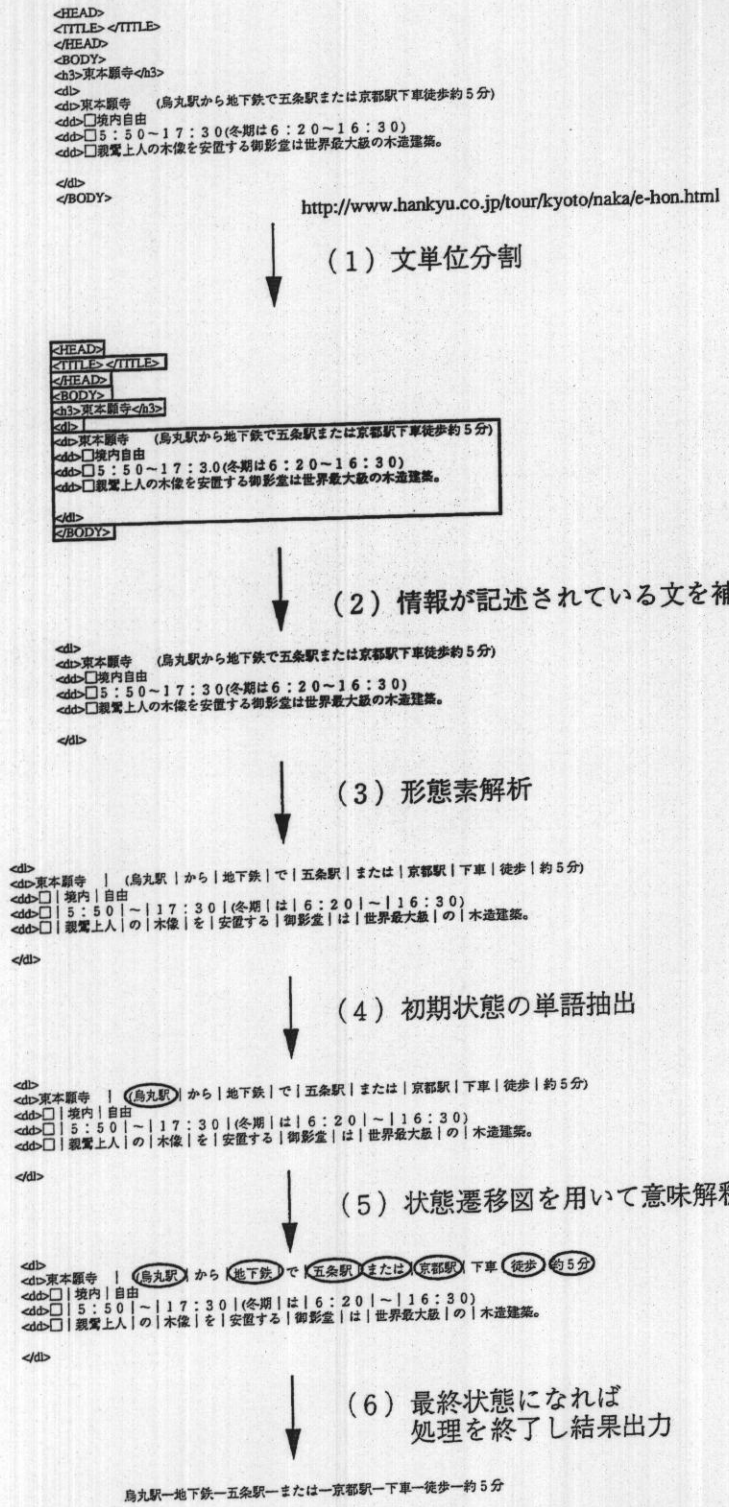


図 2.10 状態遷移図を用いる手法のフロー

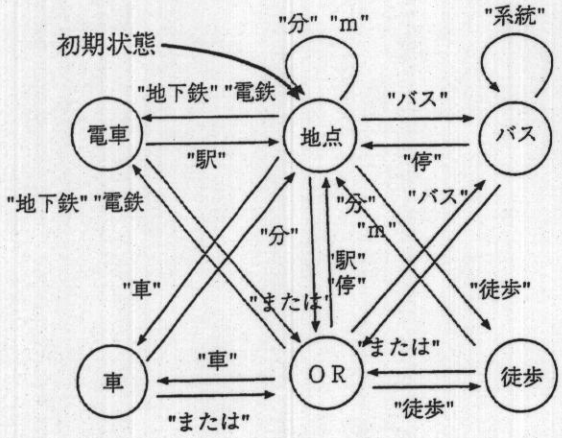


図 2.11 交通手段記述を解釈するための状態遷移図

[step5] 状態遷移図を用いて記述文の解釈処理を行なう (図??の (5))。

[step6] 解釈処理中、最終状態に達したと判断されれば解釈処理を終了する。最終状態の判定基準は、(1)10 単語連続して状態遷移が起こらない、(2) 文の終わりに達する、(3) 現在の状態から遷移する先がない、の三つである。解釈処理が終われば結果を出力する (図??の (6))。

図??の (5) を例に、step5 の解釈処理の説明をする。まず、「烏丸駅」から解釈処理が始まり、状態遷移するための単語として「地下鉄」が最初に出現する。このとき、烏丸駅から地下鉄に乗ると解釈し、{電車} の状態に遷移する。次に、状態遷移するための単語として「五条駅」が出現し、、地下鉄に乗った後は五条駅で降りると解釈し、{地点} に状態遷移する。以下同様の解釈処理を記述文の終端まで続けることによって、図??の例では「烏丸駅－地下鉄－五条駅－または－京都駅－徒歩－約 5 分」という情報が抽出される。

状態遷移図の構造　状態遷移図は図??のようになっている。特定のパターンに match する単語が出現することにより状態遷移を行なう。ここでは、状態遷移の内部構造について説明する。

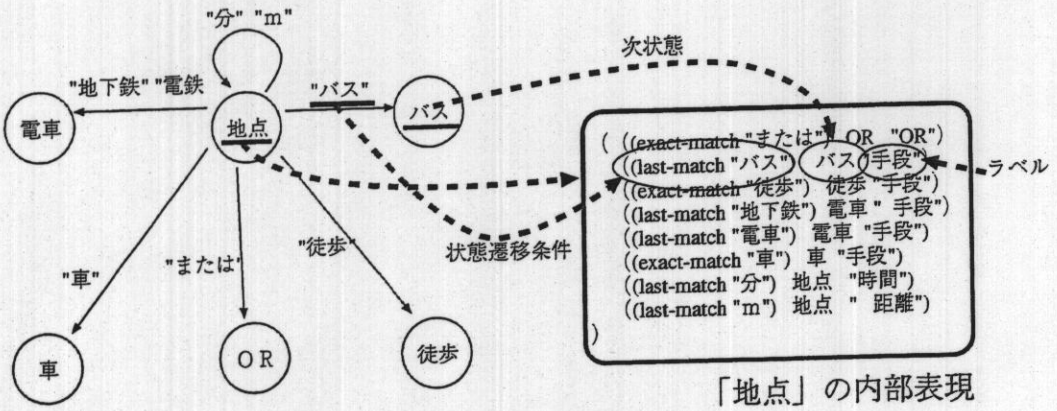


図 2.12 状態遷移図の内部表現の例

一つひとつの状態の内部構造は基本的に図?? のようなリスト構造からなっている。例えば、図??では地点の状態で「～バス」という単語が出現すれば、バスという状態に遷移し、「手段」というラベルを付ける。真になる状態遷移条件がなければ、記述文の次の単語に対して状態遷移条件の判断を行なう。表??に条件記述方式を示す。例えば「～バス」という単語パターンに一致した場合に条件遷移させたいときは、(last-match “バス”) と書く。その単語の直後に「下車」という単語がある場合に条件遷移をさせたいときは、(after x) と書く。「次状態」には次状態の名前が書かれている。次状態にも図??のような構造を記述し、次々と状態遷移を行なうようにする。「ラベル」は条件に match した単語がどのような種類であるかを示すためのものである。例えば、交通手段記述の解釈では「徒歩」「～バス」には「手段」、「～駅」「～停」には「地点」というラベルをつける。このラベルは出力する際にどのように形で出力するかを決める際に利用する。

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、情報抽出実験を行なった。

実験に使用した状態遷移図は 10 件のサンプルページを参照し、筆者 が作成した (作成した状態遷移図は図??)。表??に評価結果を示す。

表 2.12 状態遷移の条件

述語 (x は任意の文字列)	その意味
(exact-match x)	x と完全一致する単語であれば状態遷移
(partial-match x)	x を含んでいる単語であれば状態遷移
(last-match x)	単語が「～ x」であれば状態遷移
(before x)	直前の単語が x と完全一致すれば状態遷移
(after x)	直後の単語が x と完全一致すれば状態遷移
(before-partial x)	直前の単語が x を含んでいれば状態遷移
(after-partial x)	直後の単語が x を含んでいれば状態遷移
(equal h1)	<h1>に完全一致すれば状態遷移

表 2.13 状態遷移図による情報抽出実験の結果

1. 正確に記述部分が抽出されたページ数	85 ページ（100 ページ中）
2. 正確に記述部分が解析されたページ数	70 ページ （100 ページ中）

```
(defconstant 地点 '(((last-match "分") 地点 "時間")
                    ((last-match "M") 地点 "距離")
                    ((last-match "m") 地点 "距離")
                    ((last-match "バス") バス "手段")
                    ((exact-match "徒歩") 徒歩 "手段")
                    ((exact-match "車") 車 "手段")
                    ((partial-match "地下鉄") 電車 "手段")
                    ((partial-match "電鉄") 電車 "手段")))

(defconstant バス '(((last-match "停") 地点 "バス停"
                    ((last-match "系統") バス "系統")
                    ((last-match "分") 地点 "時間")
                    ((exact-match "すぐ") 地点 "時間")
                    ((after "下車") 地点 "バス停")))

(defconstant 徒歩 '(((last-match "分") 地点 "時間")
                    ((exact-match "すぐ") 地点 "時間")
                    ((last-match "M") 地点 "距離")
                    ((last-match "m") 地点 "距離")))

(defconstant 電車 '(((last-match "駅") 地点 "駅名")
                    ((after "下車") 地点 "駅名")))

(defconstant 車 '(((last-match "分") 地点 "時間")))

(defconstant ＯＲ '(((last-match "駅") 地点 "駅名")
                    ((last-match "停") 地点 "バス停")
                    ((after "下車") 地点 "動作")
                    ((last-match "バス") バス "手段")
                    ((exact-match "徒歩") 徒歩 "手段")
                    ((exact-match "車") 車 "手段")
                    ((partial-match "地下鉄") 電車 "手段")
                    ((partial-match "電鉄") 電車 "手段")))

(defconstant 終 '処理終了)
```

図 2.13 実験で使用した状態遷移図の内部表現

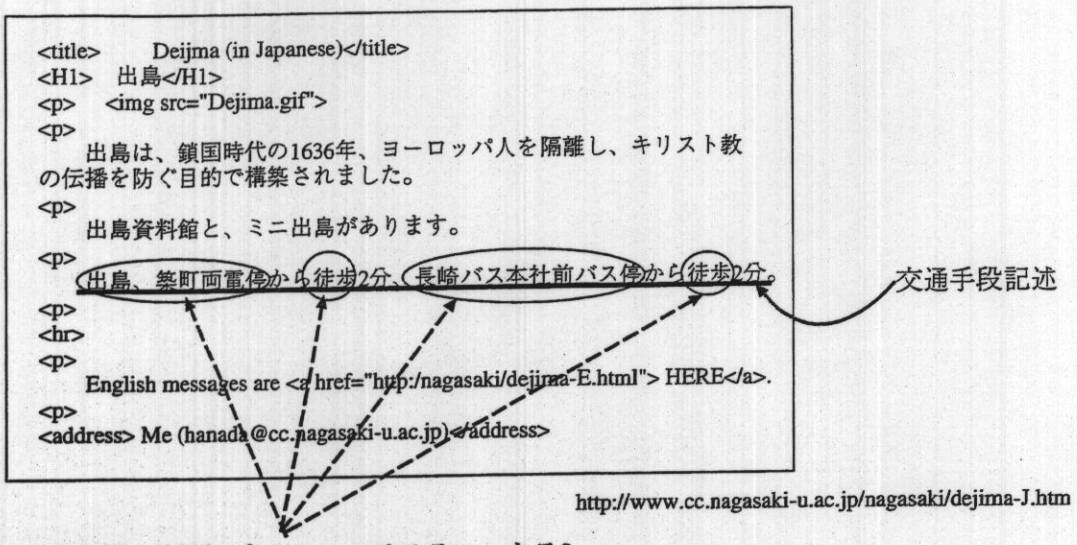


図 2.14 ヒューリスティックスが不足していた例

考察 表??の 1. では 100 ページのうち 85 ページで正確に記述部分を抽出したことを表しており、2. では 100 ページのうち 70 ページで正確に記述部分を解析したことを表している。つまり再現率を意味している。それに対して適合率は、$70/85 \times 100 = 82.3$ %である。

正確に記述部分が抽出された割合については、二つのヒューリスティックスだけを利用したにもかかわらず、85 %という比較的高い正解率が得られた。一方、残り 15 %の誤りの原因はヒューリスティックスの不足であることがわかった。

ヒューリスティックスの不足の典型例を図??に示す。この WWW ページには、交通手段が書かれていることを示すタイトルもなければ、駅名らしき記述も見られない。しかし、交通手段であることを示す記述として「〜停」や「徒歩」の単語があり、ヒューリスティックの改善によって抽出可能と思われる。

次に、正確に記述部分が抽出されたページのうちで、正確に記述部分が解析されたページの割合は $70/85 \times 100 = 82.3\%$と比較的高い率が得られた。一方、誤りの原因としては、(1) 状態遷移図による表現の限界と (2) ヒューリスティックスの不足、が考えられる。

(1) については、実験の結果、ルールにない同義語で記述されている場合、情報抽出がうまくできないケースがあり、状態遷移図を利用して解釈処理をする際には、同義語の処理については考慮する必要があることがわかった。

(2) については、状態遷移図として記述可能だが、未記述の部分がいくつか見られた。

2.6.3 概念の記述ルールを利用した手法

本研究における概念とはいくつかの要素（スロット）から構成される項目のことを指している。例えば、ここでいう概念として「温泉」について考えてみると、温泉について書かれている WWW ページには「神経痛に効能がある」や「露天風呂がある」のような記述がよく見られる。

そこで本研究では、ある分野の WWW ページによく見られる記述をルール化し、それらを該当する分野の WWW ページに適用して情報抽出を行なう手法を提案する。

次に、本手法の処理手順について具体例を挙げて説明する。

本手法の処理手順 ここでは、温泉の WWW ページを例に挙げて本手法の処理手順を説明する。温泉の WWW ページによく見られる表現として「～病に効果がある」を取り上げる。ここで、「～病に効果がある」をルール化すると、

(1)「効果」という単語が出現する文に「～病」という単語が出現すれば、
「～症に効果がある」と解釈する。

になる。このルールを用いて、WWW ページに書かれている温泉が何の症状に効くのかを抽出する手順を以下に示す（図??参照）。

[step1] まず、WWW ページを文単位に分割し、情報抽出が行なえるようにする（図??の (1)）。

[step2] ヒューリスティックスを適用できるようにするために、WWW ページ全体に対して形態素解析を行なう（図??の (2)）(ヒューリスティックスの詳細は以降で説明する)。

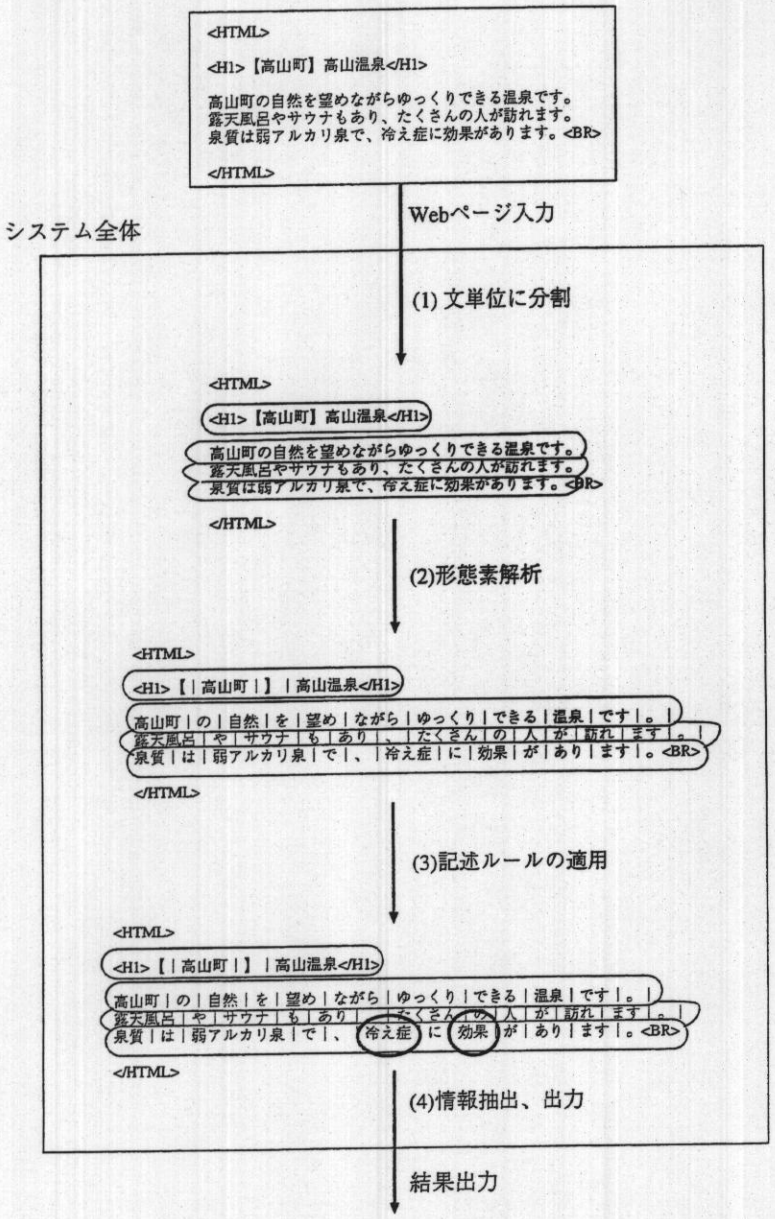


図 2.15 概念の記述ルールによる手法の処理フロー

[step3] ルールに match する文を WWW ページの最初から探索する図??では (1) に沿って、「効果」と「～症」の両方の単語を含む文を抽出している（図?? の (3)）。

[step4] ルールに match する文があれば抽出し、ルールに従ってユーザが求める情報を出力する（図??の (4)）。図??では、(1) により「～症」に効くことがわかったので、「～症」をユーザに対して出力する。

以上の処理から図??の WWW ページに書かれている温泉は「冷え症」に効果があることがわかる。

ルール化の手順　記述を抽出するルールの設定は、次のように行なう 。

最初に、概念がどのような要素から構成されているかを設定する。例えば、お寺の記述について調べるのならお寺はどのような特徴を持っているのか、つまりお寺がどのような要素 (スロット) から構成されるかを設定する。

e.g. 温泉の構成要素->温泉の名前、風呂の種類、温泉の質、温泉の効能の四つ

次に、どのような要素から構成されるのかを決めた後、各構成要素について、WWW ページ中ではどのように記述されているかを設定する。例えば、寺の創建年なら「“創建年” という単語が出現する文の中には “????年”(?は数字) という単語が記述されている」と各要素についてルールを設定する。

e.g.「温泉の効能」の記述ルール->「効能」or「効果」or「効く」
という単語がある文に
「～症」or「～痛」or「～病」が
記述されている。

ここでのルール設定は以下の手順で行なう。

[step1] 概念（「温泉」「寺」「研究室」など）を決める。

[step2] その概念にはどのような事柄が書かれるのかを考える。

表 2.14 概念の構成要素を示すための述語

述語（x は構成要素の名前）	その意味
(has-one x)	概念の構成要素として x を 0 もしくは 1 個だけ持つ
(has-some x)	概念の構成要素として x を 0 個以上持つ
(has-at-least x)	概念の構成要素として x を少なくとも 1 個以上持つ
(is-a x)	概念が x という概念を包含している （概念 x は別途定義されている）

[step3] それぞれの事柄は WWW ページではどのように記述されてるかをヒューリスティックなルールで記述する。

次節から、具体的なルール設定方法について述べる。

概念を構成する要素の設定　概念にどのような構成要素を持つかを示すための述語をいくつか用意した。表??に構成要素を示すための述語を示す。

表??を用いてどのように概念の構成要素を設定するのかをを用いて説明する (図??)。図??では「寺」という概念について設定する。寺の名前が一つあるということを設定する場合は、表??にある has-one の述語を用いて (has-one 寺の名前) と記述する。創建年も同様に has-one の述語を用いて (has-one 創建年) と記述する。国宝がいくつか書かれているということを設定する場合は、表??にある has-some の述語を用いて (has-some 国宝) と記述する。重要文化財についても国宝と同様に has-some の述語を用いて (has-some 重要文化財) と記述する。訪問地については概念として設定されているものとして、表??の is-a の述語を用いて (is-a 訪問地) と記述する。訪問地の構成要素についても寺と同様に記述する（図??参照)。

WWW ページ中での記述ルールの設定　各構成要素が WWW ページ中でどのように記述されているかを設定するために、述語をいくつか用意した。表??にその述語を示す。表??の x、y は単語パターン、述語または構成要素名である。単語パターンについては後述する。

寺->寺の名前が一つ、創建年も一つ、国宝、重要文化財が
いくつか書かれている訪問地である。
訪問地は一つの電話番号と一つ交通手段から構成されている。

↓

```
寺->      ((has-one 寺の名前)
           (has-one 創建年)
           (has-some 国宝)
           (has-some 重要文化財)
           (is-a 訪問地)
          )
訪問地-> ((has-one 電話番号)
           (has-one 交通手段)
          )
```

図 2.16 概念の構成要素の設定例

表 2.15 WWW ページ中での記述の特徴を示すための述語

述語	その意味
(is x with (y))	y が出現する文の中に出現する x。
(or x y)	x もしくは y が文中に出現する。
(and x y)	x と y が文中に出現する。
(list x y)	文中で、x の直後に y という単語が出現する。

表 2.16 単語パターンの記述方式

述語	その意味
"<温泉>"	「温泉」に exact match
"温泉"	「温泉」に partial match
"+温泉>"	「～温泉」という単語に match
"<温泉+"	「温泉～」という単語に match する単語
"*"	任意の単語に match する
">漢字<"	漢字を含んでいる単語に match
">カナ<"	カタカナを含んでいる単語に match
">かな<"	ひらがなを含んでいる単語に match
">記号<"	記号を含んでいる単語に match
">ローマ字<"	アルファベットを含んでいる単語に match
">数字<"	数字を含んでいる単語に match
"*名詞*"	品詞が「名詞」である単語に match(他の品詞でも可)
"$普通名詞$"	品詞の格が「普通名詞」である単語に match(他の品詞の格でも可)
"?食品名?"'	あらかじめ「食品名」をラベルに持つ単語群を用意しておく。 その単語群の中のどれかの単語と partial match する。

単語については部分照合も行なうため、単語パターンにもいくつか記述方式を用意した。表??にその記述方式を紹介する。

図??の構成要素の記述ルールを例を用いて、どのように記述ルールを設定するのかを説明する。

まずは単語パターンについて考えることにする。図??の「効能」、「効果」、「効く」は exact match させたいので、"<効能>"、"<効果>"、"<効く>"という単語パターンになる。同様に「～症」、「～痛」、「～病」についても表??を参考にすると、"+症>"、"+痛>"、"+病>" という単語パターンになる。

「"+症>", "+痛>", "+病>"」および「"<効能>", "<効果>", "<効く>"」については、(or x y) の述語を使い、それぞれ、(or "+症>" "+痛>" "+病>") および (or "<効能>" "<効果>" "<効く>") と記述する。

ここで、(or "+症>" "+痛>" "+病>") に対しては新たに「症状名」という名の

効能-> 「効能」または「効果」または「効く」が出現する文の中に
「～症」または「～痛」または「～病」という単語も出現する。

↓

```
効能-> (is 症状名
          with (or "<効能>" "<効果>" "<効く>"))
```

図 2.17 構成要素の記述ルールの例

構成要素を定義すれば、次のようなルールになる。

$$症状名- > (or" + 症 > "" + 痛 > "" + 病 > ")$$

記述ルールの設定としては、(or "<効能>", "<効果>", "<効く>") がある文の中に"「症状名」"がある場合に match すればいいので、(is x with (y)) を利用する。つまり、(is (or "+症>", "+痛>", "+病>") with 「症状名」) と設定すればよい。

実験 本研究ではあらかじめ分類しておいた WWW ページを、温泉について 100 ページ、飲食店について 100 ページ、寺について 100 ページ用意した。それらを対象に情報抽出処理を行なう実験をした。

実験に用いる WWW ページは以下のような手順で選出した。まず、サーチエンジン[2]で「料理」、「温泉」、「寺」をキーワードとして、WWW ページを検索した。「料理」のキーワードで検索した WWW ページの中から飲食店に関して書かれているページを 100 だけ取り上げて実験に用いた。他も同様で、「温泉」のキーワードで検索した中から 100 ページ、「寺」のキーワードで検索した中から 100 ページを取り上げて実験で用いた。

ルールの記述は、それぞれの分野に対して 10 ページのサンプルを参考に作成した (使用したルール については、付録 A1 ～ A5 を参照)。

[2] http://www.info.waseda.ac.jp/search.html

表 2.17 概念記述ルールによる情報抽出実験の結果

分野	適合率	再現率
温泉	82 %	61 %
寺	72 %	73 %
飲食店	85 %	41 %
3 分野の平均	79 %	58 %

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、それぞれついて分野ごとの平均と全分野での平均を計算した。その結果を表??に示す (実験の典型例は図??参照)。

再現率＝ (正しく抽出した *item* 数/本来抽出すべき *item* 数) × 100

適合率＝ (正しく抽出 *item* 数/実際に抽出した *item* 数) × 100

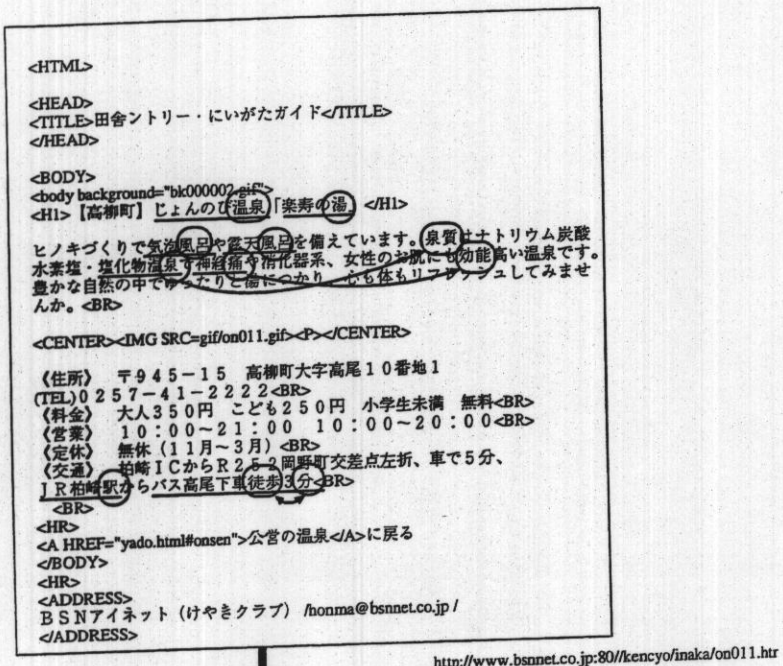
考察 実験の結果、簡単なヒューリスティックスを使っただけにも関わらず比較的高い精度で情報が抽出できている。

表??の結果を 3 分野平均で見ると、再現率が 6 割、適合率が 8 割であるが、飲食店の再現率が極端に低い。その原因は、料理名に対する知識のルール記述が不足していたことにある。つまり、飲食店のページには料理名が多く見られるが、本研究で用いた知識では全ての料理名を網羅できなかった。また、本研究で構築したシステムでは知識獲得ができないので、知識が不足している場合に対処できない、という短所がある。

2.7 第 2 章のまとめ

本章では、オントロジーを用いた、情報の収集・分類・抽出法を提案し、インターネットからの情報収集・整理システム IICA を実装した。また、WWW を対象に、IICA の各方法の評価を行なった。これらの結果から、本章で提案したアプローチには次のメリットがあることがわかった。

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。



```
<HTML>

<HEAD>
<TITLE>田舎ントリー・にいがたガイド</TITLE>
</HEAD>

<BODY>
<body background="bk000002.gif">
<H1>【高柳町】じょんのび温泉「楽寿の湯」</H1>

ヒノキづくりで気泡風呂や露天風呂を備えています。泉質はナトリウム炭酸
水素塩・塩化物温泉で神経痛や消化器系、女性のお肌にも効能高い温泉です。
豊かな自然の中でゆったりと湯につかり、心も体もリフレッシュしてみませ
んか。<BR>

<CENTER><IMG SRC=gif/on011.gif><P></CENTER>

〈住所〉　〒９４５－１５　高柳町大字高尾１０番地１
(TEL)０２５７－４１－２２２２<BR>
〈料金〉　大人３５０円　こども２５０円　小学生未満　無料<BR>
〈営業〉　１０：００～２１：００　１０：００～２０：００<BR>
〈定休〉　無休（１１月～３月）<BR>
〈交通〉　柏崎ＩＣからＲ２５２岡野町交差点左折、車で５分、
ＪＲ柏崎駅からバス高尾下車徒歩３分<BR>
<BR>
<HR>
<A HREF="yado.html#onsen">公営の温泉</A>に戻る
</BODY>
<HR>
<ADDRESS>
ＢＳＮアイネット（けやきクラブ）/nonma@bsnnet.co.jp/
</ADDRESS>
```

http://www.bsnnet.co.jp:80//kencyo/inaka/on011.htr

温泉名：じょんのび温泉
風呂の名前：楽寿の湯、気泡風呂、露天風呂
泉質：塩化物温泉
効能：神経痛
最寄り駅：ＪＲ柏崎駅
アクセス方法：徒歩３分

図 2.18 概念の記述ルールを利用した手法の実験例

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ー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。

第3章

オントロジーを利用した現場技術情報の共有

3.1 はじめに

第2章では、WWWに代表されるような多様性(形式面、内容面)、分散性、大規模性を扱う情報源群を対象とした、オントロジーによる情報収集・分類・抽出法について議論を行なった。

本章では、より専門的な背景知識が必要となる工学的問題を取扱う。ここでの目的は、企業などの特定された実務的な目的を持つ組織の中において、様々な業務・技術に関する情報を共有化を実現するための方法論について検討し、明らかにすることである。

本研究では、ICoB:(工学的知識の体系化と共有のための知的ドキュメントベース)の枠組に基づき、対象に関するオントロジーを用いて、ベテラン技術者の専門知識やノウハウの継承を促進する現場技術情報共有支援法を提案する。 例として「業務マニュアルが整備されている」業務と「マニュアル化されていない」業務とが渾然一体となっている「配電用変電所における流用変圧器の改修計画業務」支援を対象に選定し、現場技術情報共有支援システム OnTheSpot を実装し、現場のベテラン技術者によるフィールドテストを行ない、本システムの有効性について検証する。

本章では以下の通り構成されている。3.2節では、現場技術情報の定義を行ない、組織化・共有化における問題点を明らかにする。3.3節では、現場技術情報の組織化・共有化のためのアプローチとして、ICoB:を提案する。3.4節では、応用例として、ICoBアーキテクチャに基づく変圧器改修計画業務支援システム OnTheSpot

について紹介する。3.5節では、オントロジーを用いた知的検索法について紹介する。3.6節では、本アプローチついて評価・考察を行なう。

3.2 現場技術情報の共有の現状

技術者・作業者は、業務を遂行する過程でさまざまな技術情報を参照する必要がある。企業内では、円滑な業務遂行のために基準・規定類（以下、業務マニュアル）が整備されている。また、近年ネットワーク技術の進歩により、これらの書類は電子化され、イントラネット (Intranet) 上に蓄積されるようなってきた。

しかし、これらの技術情報の利用方法は必ずしも自明ではなく、職場への新規配属者にとっては、膨大な量の書類の中から当該業務を遂行する上で必要な業務マニュアル内容を見出すことは容易ではない。さらに、非定型な業務の場合には、この業務マニュアルに書ききれない各現場業務固有の情報も多く存在する。したがって、技術情報を利用しながら業務を行なうためには、経験やノウハウが必要であり、職場への新規配属者が円滑に業務を遂行できるようになるために長期間の準備期間を要している。すなわち、単に情報が電子的に共有されているだけでは不十分であり、その使い方も含めた知識の共有が必要である。

3.2.1 現場技術情報の種類

現場の技術者が利用する技術情報は、(1) 業務手順書、管理規則、ユーザマニュアル、研修用教材、設計仕様書、規格表、入門書など一般化・マニュアル化された情報と (2) 空間的・地理的情報、色、音など現場固有の情報、(3) (1)、(2) を結びつける経験的知識やノウハウなどマニュアル化されていない情報がある。

ベテラン技術者は、自分の経験や知識も基づいて状況に応じてそれらの情報を使い分け、業務を遂行している。これらの技術情報は、表??のように、手順、構造、概念、事実、原則に関する5種類の基本情報に分けることができる。

知識の継承を円滑するためには、経験のない若手技術者はまず業務に関する手順を把握するとともに、ベテラン技術者が持つ対象に関する構造レベルや概念レベルでの知識を充分理解しなければならない。そして、それらに基づき、現場の状況で得られるデータなどの事実や様々な制約条件となる原則に従いながら、業

表 3.1 現場技術情報の種類

種類	内容
1. 手順	特定の結果を得るために実行する一連の作業に関する知識. 作業に先立って行なう判断や判断の結果に基づく動作も含む。
2. 構造	物理的な対象や複数の要素に分解できる対象の、見え方や構成に関する関する知識。図面、現場の3次元画像なども含む。
3. 概念	用語、アイデア、抽象的な概念などの意味に関する知識。
4. 事実	現場で観測して得られるデータや過去の事例など。
5. 原則	遵守・禁止事項に関する知識。規則、方針、ガイドライン、警告、注意、通則、定理、仮定、原理、前提条件なども含む。

務を進めることになる。

したがって、現場技術の共有では、業務の手順を示すとともに、若手技術者が判断や処置にとまどう事項や忘れてしまった事項を、対象の構造レベルや概念レベルでの知的支援が有効であると考えられる。

3.2.2 従来のアプローチの限界

近年、数多くの業務支援システムがエキスパートシステムとして構築され、業務の効率化・信頼性向上の面で成果が期待されてきたが、知識の獲得と保守が高コストとなりそれほど普及してこなかった。また、従来のシステムでは、非定型でマニュアル化されていない対象固有のノウハウをうまくルール化して扱うことが困難なため、実用に耐え得る業務支援が行えなかった。

これは、「業務マニュアルが整備されている」技術情報と「マニュアル化されていない」技術情報を渾然一体に扱い、個々人のノウハウとして蓄積されていた知識を随時簡単にシステムに組み込み、皆で共有して利用していくことが困難であったことが、大きな要因であると考えられる。例えば、コンピュータには詳し

くないが、個々の業務遂行ノウハウを豊富に保有する人がワープロ感覚で自由に知識をシステムに組み込むといったことは困難であり、通常は、ある形式で知識を整理し特定の形式で入力する必要があり、個々人の知識を自由に入力出来る状況でなかった。

知識処理の専門家でない現場の技術者が自由に情報基盤を利用することができるためには、データベース・ビューや情報検索といったシンタックス的処理を中心とする情報技術から、日常的な常識レベルの知識、個々人が持つ断片的な知識やアイデアなどの情報を表現し、内容レベルでの共有化、知識の拡大、コミュニケーションを実現するための内容指向のアプローチが必要になる。

3.2.3 現場技術情報共有における課題

現場技術情報共有を実現するためには、次の課題をクリアする必要がある。

1. 情報収集・整理・理解のボトルネック: 技術者が、大量の技術情報の中から必要なものをだけを収集・整理し、理解するには時間・労力を要する。

2. 情報利用者と情報提供者の分離: 情報利用者と情報提供者か完全に分離されている場合、現場で得られた情報を他の技術者のために蓄積・提供できない。

3. メンテナンスの問題: 状況や規格の変更に伴うドキュメントの追加・変更や整合性のチェックを人手だけで行なうのは困難である。

4. 現場でしか得られない情報の取扱い: 机上ではなく、作業現場いる技術者が3次元的に認識してはじめて得られる情報が、すでに蓄積されている技術情報のどれに該当し、利用可能どうか判断することは難しい。

3.3 オントロジーに基づくドキュメント処理

本節では工学的知識の性質を考察して、この工学的知識の処理としてオントロジーに基づくドキュメント処理を提案する。

3.3.1 工学的知識の様相

設計、改修、メインテナンスなどにおいて用いられる知識、すなわち技術者がもつ工学的知識はその形態においても、性質においても多種多様であり、これが従来のエキスパートシステムが必ずしも成功しなかった原因である。ここでは主に表現形態と共有性の二つの視点から工学的知識を分類する。まず、表現形態としては、モデルレベル、形式化レベル、メディアレベルの3つを考えることができる。モデルレベルとは対象をその振る舞いを予測可能なレベルで表現することであり、多くは数式で表現される。形式化レベルにおいては、抽象的な概念などで対象は記述される。ここでは振る舞いの予測は可能ではないが、概念の操作により、多様な表現が生成可能である。メディアレベルでは、対象はその表現に用いられたメディア（自然言語、画像など）によってのみ記述される。ほとんどの工学的問題では、この3階層のレベルいずれかの1つの階層レベルのみを使って対象を表現することはできず、3つのレベルの知識を密に結合して持つことで、多様な工学的問題解決が可能となる。

また共有性とはその知識がどのような範囲において共有されるものであるかという視点である。個人知、共有知、世界知のようにわけることができる[?]。常識的知識は世界知であるが、あるグループ（例えば会社内）で共有されるものは共有知である。さらに個人的に貯えれられた個人知である。特に技術者は個人知を交換しながら自らの個人知を拡充したり、共有知を生成させたりしていく。

3.3.2 ICoB：オントロジーに基づくドキュメントベース

ここでは、オントロジーをこれらの表現形態の統合と個人知と共有知の統合の手段として用いる。オントロジーについては多様な意味づけが存在するが[?]、ここでいうオントロジーとは「物事を表現する際の基本となる単位(プリミティブ)の体系」とする[3]。

オントロジーは表現形態としては形式化レベルに含まれ、共有性という意味では共有知に含まれる。すなわち、オントロジーは多様なメディアレベルの表現や

[3]すなわちここでは、溝口[?]の中で述べられているレベル3に相当する程度までの能力をオントロジーに求めている。

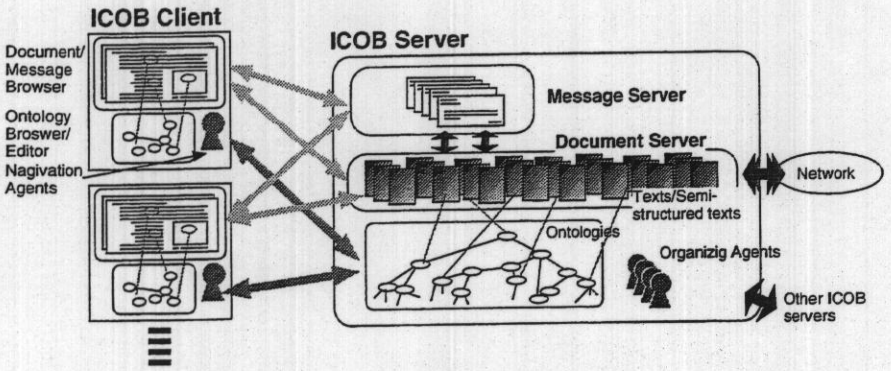


図 3.1 ICoB アーキテクチャ

モデルレベルの表現をつなぎあわせる。また、同時に個人やグループの知識をつなぎあわせる役割も行なう。ここでは以上の考えに基づき、オントロジーに基づくドキュメントベース ICoB アーキテクチャを提案する (ICoB のアーキテクチャは図??を参照)

ここで、ドキュメントと呼ぶものは個人および集団でもつ書類や図面、メッセージを指す。これはメディアレベルでの知識の表現である。このドキュメント群を管理するシステムをドキュメントベースと呼ぶ。オントロジーに基づくドキュメントベース ICoB においては、ドキュメントはオントロジーの中の概念に結びつけられて管理される。すなわち、ドキュメントはメディアの種類や利用するアプリケーションなどで管理されるのではなく、その内容（知識）によって管理される。ユーザはオントロジー中の概念を組み合わせることで、必要なドキュメントを表現する。また、必要であればユーザはオントロジーを追加・変更することでドキュメントベースを更新する。

ICoB には次の 3 つの特色がある。

1. ユーザや状況に応じた表示や利用方法の提供: 技術者の知識・経験や視点に応じて、技術情報を分かりやすく整理し、提示する。クライアント、サーバーはユーザの情報組織化を支援するための分類・抽出・統合化などの機能を持つ。

2. 情報利用者と情報提供者の融合: 技術者は、共有されたオントロジーや個人用オントロジーを用いることでドキュメントやメッセージの検索や提出を行ない、業務によって得た知識を他の技術と共有することができる。

3. メンテナンスの支援: 知識処理の専門家でなくても、容易に知識を入力し、編集することができる。また、規格の変更に伴う大量のドキュメントの追加・変更や整合性のチェックも、オントロジーに基づく関連項目の分類・構造化によって行なう。

3.4 OnTheSpot:現場技術共有支援システム

具体的開発対象として「業務マニュアルが整備されている」業務と「マニュアル化されていない」業務とが渾然一体となっている「配電用変電所 (図??参照) における流用変圧器の改修計画業務」支援を対象に選定した。この改修計画業務においては、変圧器が標準に保有すべき機能は規格類で詳細に規定されているが、改修対象の変圧器が流用品であるゆえに古く、標準工法で所要機能を整備することができない場合が多く、個々の変圧器に特有な工法を考案する事が必要となる。このため、従来の方法では業務支援のシステム化が出来なかったものである。

変圧器改修計画業務の支援では、新規担当の改修設計者が、変圧器の改修設計を実施する際に、変圧器のどの部分が、どのような状態のときに、どう改修すればよいか迷いやすい事項・内容について、判断を支援することが主要な課題となる。

3.4.1 開発方針

このような業務支援システムを構築するためには、技術者が改修設計業務においてどのような情報をどのように利用しているか明らかにするとともに、マニュアル化されていないベテラン技術者が持っている経験的知識を再利用可能な形で記述・整理する必要がある。そして、マニュアル化された情報とマニュアル化されていない経験的知識や現場の情報とをうまく結びつけて利用できるようにしなければならない。

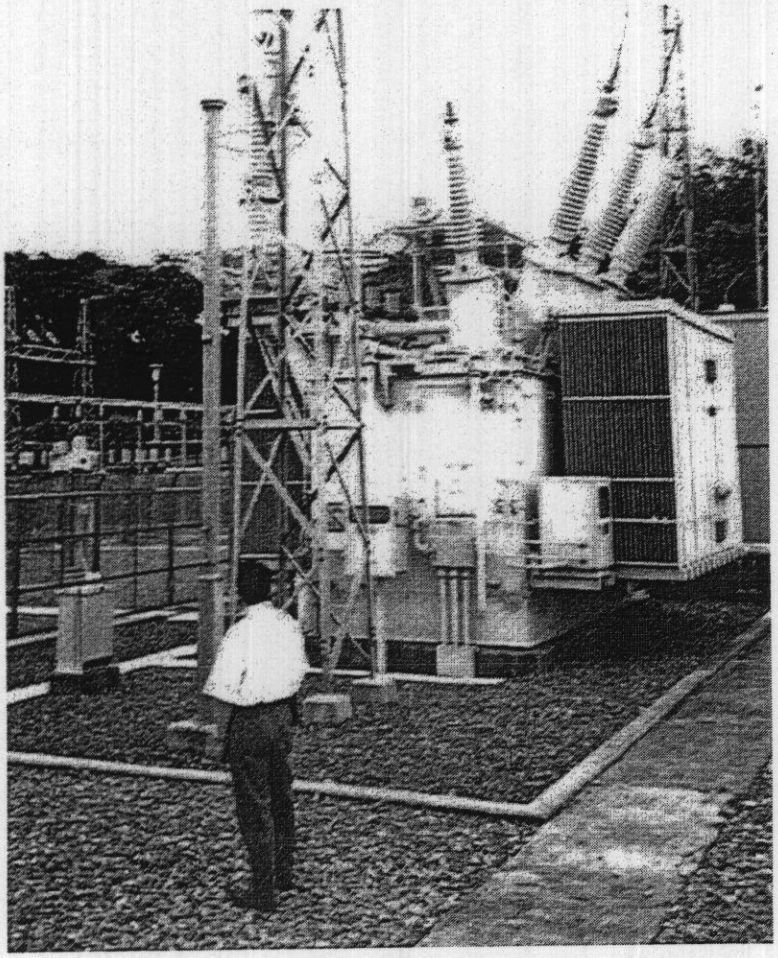


図 3.2 配電用変圧器

ここでは、ICoB の考え方に基づき、変圧器改修計画業務知識のドキュメントベース化を行った。

本システムの基本的動作は、現場の状況を入力することで改修項目を提案するというものである。このような改修計画業務を支援するという意味では、エキスパートシステムの一種であるが、ここではドキュメントベースの一つの機能として実現している。すなわち、ドキュメントは改修支援だけでなく、関連項目の検索などの他の用途に用いられる。

本システム開発にあたって次のような要求を満たすことを目標とした。

1. ユーザや状況に応じた表示や利用方法の提供: 技術者の知識・経験や視点に応じて、技術情報を分かりやすく整理し、提示する。
2. 容易な知識入力: 知識処理の専門家でなくても、容易に知識を入力し、編集することができる。

3.4.2 システム構築手順

システム開発に当たり、まずノウハウのもととなる資料を現場の専門家 (ベテラン技術者) に作成してもらい、それをもとに以下の手順でシステムを構築した。

1. 専門家による改修業務計画手順の資料作成
2. HTML 形式への自動変換 (ハイパーテキスト化)
3. 重要語句と静止画・動画・資料データ間のリンク付け

以下では、各プロセスについて詳しく説明する。

改修業務計画手順の資料作成　改修設計計画業務おいて、若手技術者が判断や処置にとまどうと思われる事項について、現場の専門家が、A4 版で約 80 ページ分の資料を作成した。

この資料では、図??の上部の「ベテラン技術者が作成したドキュメントの一部」および「改修検討項目」の枠内に示されているように、(1) 改修検討項目 (大項目

14、中項目約 61、小項目約 140) と、それに対応する (2) 改修検討フローおよび (3) 改修対策案について、インデントや矢印などを用いて記述されている。したがって、知識工学の専門家ではない現場の技術者でも、ワープロで文章を書くのに近い感覚で、現場のノウハウや知識を記述することができる。

ドキュメントのハイパーテキスト化　ベテラン技術者が作成した資料を、各検討項目、検討フロー、対策案単位のドキュメントに分割化し、各ドキュメントを HTML 形式に変換することにより、専門家の改修計画業務をハイパーテキスト化する。図??は、LR 制御箱の覗窓に関する知識をハイパーテキスト化した例である。

この作業は、資料の持つ弱い構造を利用したオーサリングツールによって自動的に行なう。これは実際には資料を検討フローの各段階として分割し、各段階を 1 つのハイパーテキストに、次の段階への指示をその中のハイパーリンクとして付加するという形で実行される。

重要語句と静止画・動画・資料データ間のリンク付け　変圧器の改修設計者が利用する技術情報には、業務手順書、管理規則、 設計仕様書、用品規格、入門書など様々なものが存在する。しかし、実際の業務では、ベテラン技術者は、自分の経験的な知識に基づいて、現場の状況に応じて判断し、情報を使い分け、業務を遂行している。また、実際の業務では、教科書やマニュアルにはない、色、音、空間的な広がりなどの現場で経験しないと得られない情報も非常に重要である。そこで、現場技術マニュアル、用品規格、教科書などのマニュアル化された知識や現場の静止画像、動画像等をドキュメントベースに取り込むことによって、改修計画業務のメディアレベルでの理解支援を行なう。

各種のマニュアル的資料や現場の静止画、動画は、専門家がリストアップした重要語句をもとにインデックス化されており、オーサリングツールによって、前項で作成したドキュメントベース内の各重要語句に自動的にリンク付けが行われる (図??参照)。また、これらの重要語句は、次節で説明する変圧器に関するオントロジーの語彙として用いられており、形式レベルの知識とメディアレベルの知識を結びつける役割を果たしている。

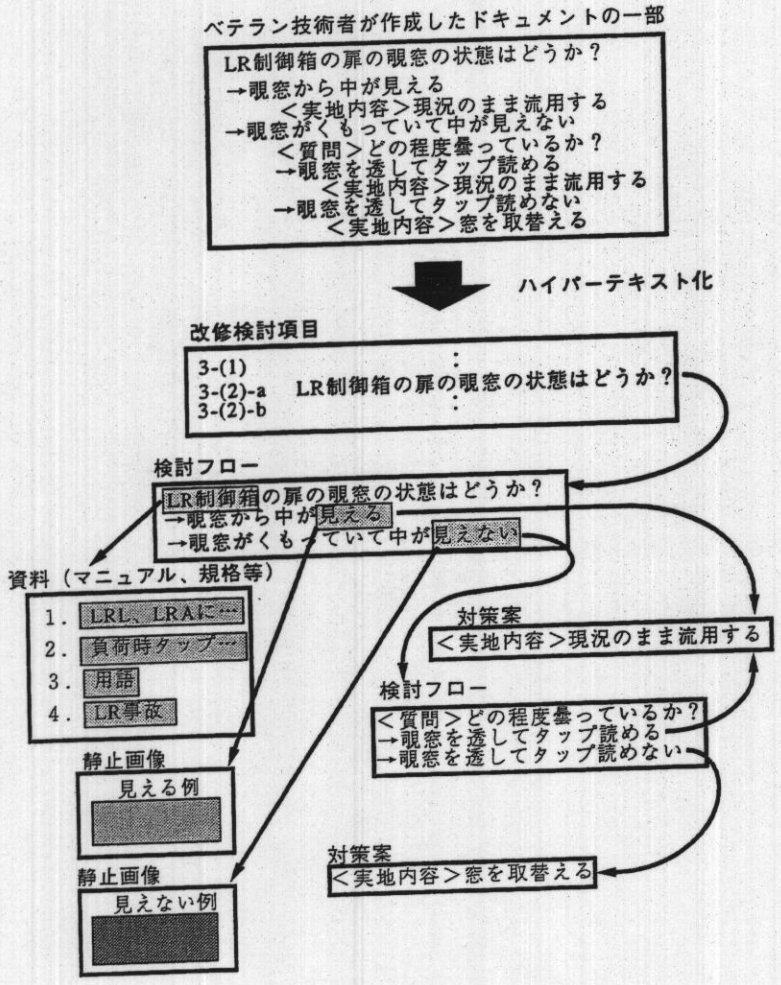


図 3.3 専門知識のドキュメントベース化

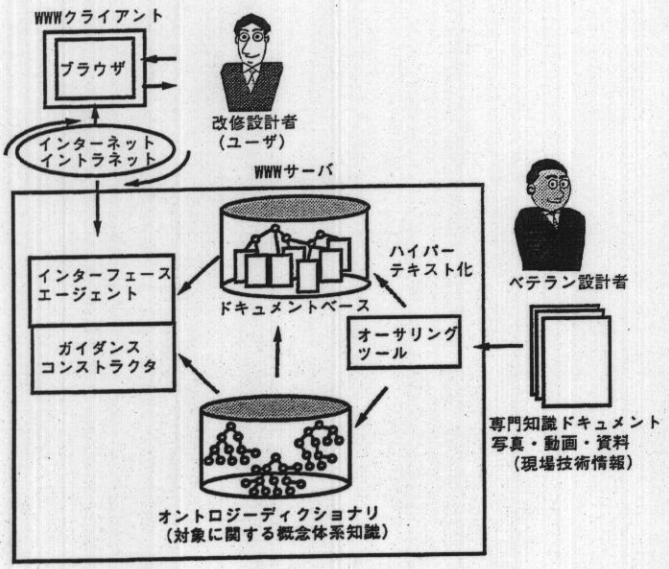


図 3.4 システムの構成

3.4.3 システムの実際

ここでは、前章までで述べた手法に基づいて実装した改修業務設計支援システム OnTheSpot のプロトタイプの概要、各機能について述べる。

システム構成 システム構成を図??に示す。本システムのユーザとなる改修設計者は、図??に示すように、パソコン端末上の WWW ブラウザを用いて、画面に表示される質問に答えていけば、変圧器改修計画書が自動的に作成される。その途中で調べたい事項が発生した場合、画面上の関連用語を選択すれば、現場技術ガイドや静止画・動画等の技術情報が表示される (図??参照)。

関連ソフトウェアはサーバのみにあり、クライアントは WWW ブラウザのみで利用可能なため、改修設計者は、このシステムを用いればイントラネットに接続しているパソコン (端末) からいつでも、どこからでも利用できる。

システムの機能と利用法 ここでは、システムの主要な機能について、利用手順に沿って説明する。

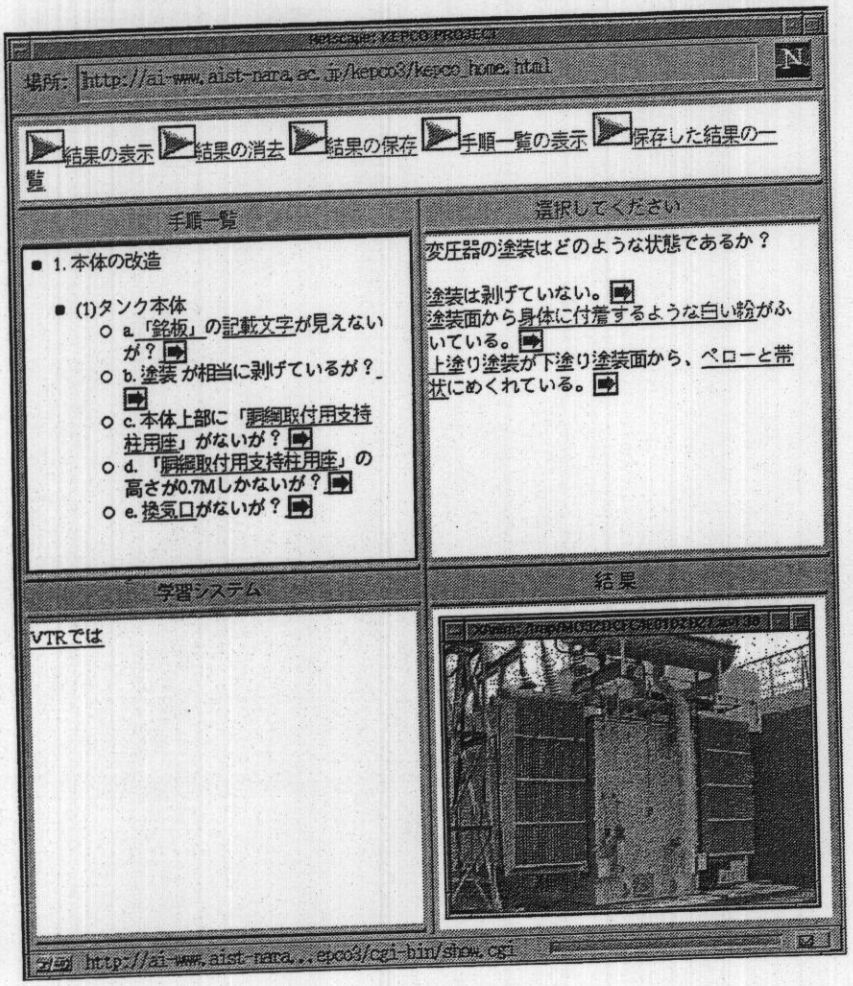


図 3.5 プロトタイプシステム

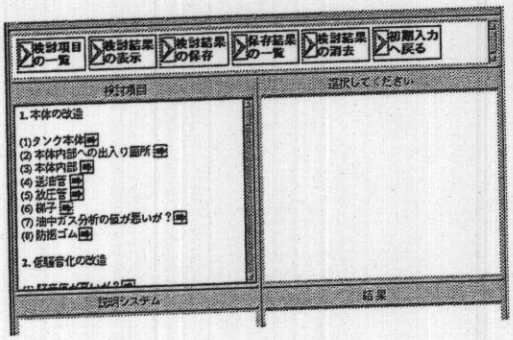


図 3.6 改修項目の一覧 (大項目)

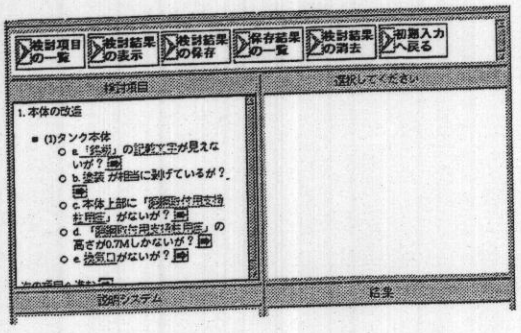


図 3.7 改修項目の一覧 (中・小項目)

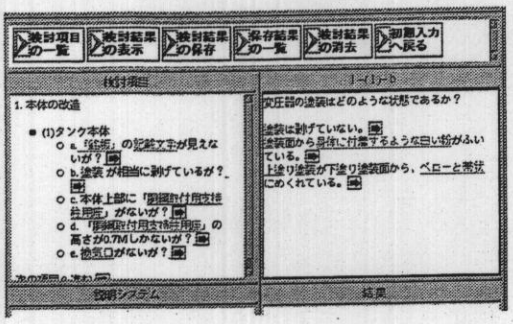


図 3.8 検討フロー

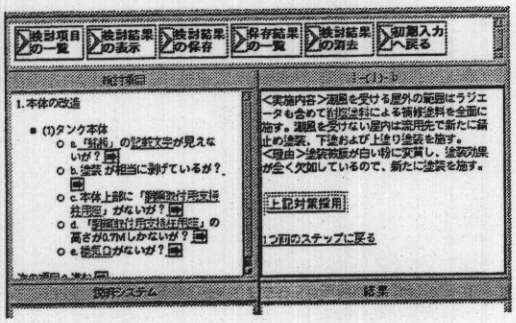


図 3.9 対策の採用

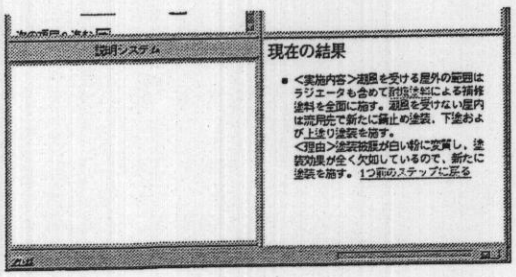


図 3.10 検討結果の表示

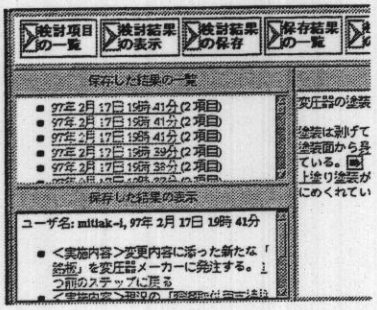


図 3.11 検討結果の保存と一覧表示

(1) 初期入力　まず、改修計画立案に必要な初期情報を入力する。初期入力画面は、タイトル画面に続いて表示される。入力内容は、工事件名、改修設計者の作業情報、流用後の変圧器の騒音仕様などの改修後の情報、流用前の変圧器の帳簿価格などの改修前の情報である。この入力データにより、騒音仕様の見直しや経年劣化によるコスト計算などをシステムが支援することが可能となる。

(2) 操作メニューの表示　初期入力が完了すると、操作画面の上部に操作ボタンが配置される。操作ボタンには、「検討項目の一覧」、「検討結果の表示」、「検討結果の保存」、「保存結果の一覧」、「検討結果の消去」、「初期入力へ戻る」がある。

(3) 検討項目の選択　「操作メニュー」の「検討項目の一覧」ボタンを選択すると、画面左上に検討項目が表示される (図??参照)。ユーザは、表示されたリスト中から項目を一つ選択する。検討項目は、大項目と中項目 (図??参照)、大項目のみ、全項目の表示 (図??参照) があり、改修設計者が選択できる。

(4) 改修検討のフロー　選択した改修項目に対応する検討事項が画面右上に質問形式で表示される (図??参照)。ユーザにはその質問の最後に改修実施事項が表示されるので、画面内の「上記対策採用」ボタンを押すと、改修実施内容の入力が完了する (図??参照)。

(5) 検討結果の表示と消去　各改修項目において改修検討のフローで決定した実施内容が画面右下に表示される (図??参照)。また、操作メニューの「検討結果の消去」ボタンを押せば今まで決定した実施内容が消去される。

(6) 検討結果の保存と保存結果の表示 (3) ～ (5) の操作を繰り返し、各項目に対する改修案が決定したら、選択メニューの「検討結果の保存」ボタンによって、これまでの検討結果をシステムに記録する、また、「保存結果の一覧」をボタンを押せば、保存した日時が画面左上に表示され、その日時を選択すると、保存した検討結果も表示される (図??参照)。

また、上記以外に、ユーザのドキュメント理解を支援する機能として、説明システムが用意されている。このシステムは、文書中の専門用語を選択すると起動され、画面左下にメニューが現れる。説明システム画面のメニューでは、選択された専門用語に対して、用品規格、現場技術ガイド、静止画像(写真)、動画像などの参照情報が項目別に表示される。調べたい項目を選択するとそれに対応する内容が表示される(図??参照)。

3.5 オントロジーの作成と利用

3.5.1 オントロジーの作成

業務支援システムにおけるオントロジーの役割は、

1. ベテラン改修設計者が持つ深いレベルの業務知識を明示的、体系的に記述し、後輩技術と共有・継承を図るとともに、
2. 現場でしか得られない情報と過去にドキュメントとして蓄積されている情報の橋渡しをする、

ことである。

オントロジーの構築では、記述する知識の粒度が問題となる。変圧器の改修設計を支援する場合、詳細な設計図面から作成したオントロジーでは、改修設計には重要でない箇所まで記述されているため、若手技術者には詳しすぎ、変圧器のどの部分が改修に重要なのかわかりくくなる。一方、単純化されすぎたモデルをもとに作成したオントロジーの場合、現場での業務には役に立たない。

そこで、専門家が作成したドキュメントに含まれている変圧器の構造に関する概念、変圧器障害の症状に関する概念、改修設計行為に関する概念をそれぞれ抽出し、これらを結びつけることでオントロジーを作成する。今回は変圧器の構造に関する概念については専門家が、それ以外はシステム開発者が行なった。

対象の構造に関するオントロジー　変圧器の構造に関するオントロジーは、改修の検討対象となる部品名、プロトタイプの説明システムで用いられている変圧器

```
(unit-026 (概念名 "ブッシング")
          (同義語 "耐塩ブッシング")
          (サブ概念 nil)
          (属性名 nil)
          (コンポーネント ("碍管" nil) ("油面計" nil) ("ＢＣ
T" nil))
          (接続関係 ("本体" "取付け架台")))
```

図 3.12 変圧器の構造に関するオントロジーの記述例

の専門用語 (ベテラン技術者が選定) 、さらに検討フローおよび改修対策案から抽出した用語など合計約 180 の基本語彙から構築した。

オントロジーは、変圧器の構造的属性に関するものとして、概念名、類義語、サブ概念、属性名、コンポーネント、接続関係について、Ontolingua[?] などのオントロジー表現言語と相互変換可能なフレーム形式で記述している (図??参照)。

改修行為に関するオントロジー ベテラン技術が作成したドキュメントでは、約 40 種類の対策行為を表現する動詞が用いられている (表??参照) 。これらの対策は、大別すると (1) 改修行為そのものに関するメタな知識 (例:「計画を見直す」、「改修しない」等)、(2) 対象の構造的属性や位置的属性を操作するための知識 (例:「～を移設する」、「～を取替える」等)、(3) 対象の表面や内部状態など構造的属性や位置的属性の変化を伴わない操作に関する知識 (例:「～を施す」、「～を洗う」等) に大別することができる。

対象の症状に関するオントロジー ベテラン技術者が改修設計にあたり対象 (変圧器) のどの部分に着目し、どのような状態を検討すべき事項としてとらえているかを客観的に把握するために、対象の症状に関するオントロジーを作成した (表??参照)。オントロジーは、ベテラン技術者が作成した改修設計に関する検討項

表 3.2 変圧器の改修設計行為に関するオントロジー

改修設計行為に関する知識の種類	専門家が用いる語彙知識
1. 改修行為に関するメタな知識	「計画を見直す」、「～を使いきる」、「～を改修しない」、「現況のままとする」、「～を止める」、「～を発注する」、「～を手配する」、「～を充当する」
2. 対象の位置・構造の変化に関わる操作	「～を移設する」、「～を設置する」、「～を設置しない」、「～を取付ける」、「～を取り外す」、「～を溶接で継ぎ足す」、「～を引きおろす」、「～構造にする」、「～を取替える」、「～を取替えない」、「～を交換する」、「～を変更する」、「～を変える」、「新たな電線にする」
3. 対象の状態の変化に関わる操作	「～を (補修) 塗装する」、「～を塗る」、「～を施す」、「グリス・アップする」、「を洗う」、「～に注水する」、「～を取り除く」、「～を拭き落とす」、「～を貼る」、「～を減らす」、「～を刻印字する」

目の中から対象の症状・状態を表現する語彙を抽出・整理して作成した。変圧器の改修設計の場合、検討の対象となる症状 (約 140 項目) は、視覚的に捕えられる情報、聴覚的に捕えられる情報に大別できる。さらに、視覚的に捕えられる情報は、対象の (1) 可視性 (例:「～が見えない」、「～が見にくい」等)、(2) 有無 (例:「～がある」、「～が設置されていない」等)、(3) 変化・変質 (例:「～が生じている」、「～が変色している」等) 、(4) 位置・距離 (例:「～の方向を向いている」、「～の裏側になる」等)、(5) 形式・型番 (例:「～型になっている」、「～に適合しない」等)、(6) 量 (例:「～が高い」、「～が多い」等) に分けて扱うことができる。

3.5.2 オントロジーによる知的検索

ドキュメントベースに格納した設計改修に関する専門知識には、通常一つの検討項目に対し複数の対策案が存在し、各対策案に到達する検討フローも複数存在する。しかし、ユーザはシステムを利用する際、各改修項目の質問に逐次回答しながら検討を進めるため、各検討フローの関連性を把握することが困難である。また、ある改修項目についての知識を追加・変更することで、他の関連する部品についての知識も変更する必要がある場合、大量のドキュメント群の中から関連

表 3.3 変圧器の症状に関するオントロジー

症状に関する知識の種類	専門家が用いる語彙知識
1. 対象の可視性	「～が見えない」、「～が見にくい」、「～を目視確認ができない」
2. 対象の有無	「～がある」、「～が無い」、「～が設置されていない」、「～が要る」
3. 対象の変化・変質	「～が変色している」、「～が生じている」「～が硬質化している」、「～から油漏れしている」、「～から油がにじんでいる」「～が欠陥している」、「～は (を) どうする」、「が (で) 汚れているが」、「～になっている」、「～が剥げている」、「～の恐れがある」、「～してある」
4. 対象の位置・距離	「～の方向を向いている」、「～の裏側になる」「～の高さにない」、「～の絶遠距離 (は良いか)」、「～へ通れない」、「～から外れている」「～がずれている」、
5. 対象の形式・型番	「～型になっている」、「～式であるが」「～に適合しない」、「～と違う」、「～が使用されている」、「～でない」
6. 対象の量	「～が高い」「～が低い」 「～を満足しない」「～が不足している」「～が足りない」「～が多い」、「～の値が悪い」、「～に近い」
7. 対象からの音	「～から異音がする」

性の高いものだけを選び出す作業は非常に労力を要する。

一方、変圧器には、異なる目的や機能をもつ変圧器の部品間でも類似の構造や原理をもつ部品が存在し、それらには似た障害の症状が現れ、ベテラン技術者が施す処置も類似している場合が多い。例えば、ブッシングの油面計と活線浄油機の温度計は、目的も機能も全く異なる部品であるが、いずれもガラス部が存在し、そこから油洩れが生じた場合、ベテラン技術者は O-リングを取り替えるという同一の対策をとる。

そこで、変圧器の構造、症状、改修設計行為に関する 3 種類のオントロジーを用いて、検討フローを分類・検索可能にし、ドキュメントの内容理解やメンテナンスを支援する方法を提案する。

3.5.3 オントロジーを利用した検討フローの分類

ドキュメントベースでは、検討項目、質問、対策案がそれぞれ独立したドキュメントとして格納されているが、ここでは、検討フロー単位で分類・検索可能にするために、1 つの検討フロー構成するドキュメント群をひとつのドキュメントとして扱う。

オントロジーを用いたドキュメントの分類法は、2.5 節の考え方に基づいている。すなわち、ドキュメントの特徴ベクトルの作成に、オントロジーの各概念 (語彙) を用いるいるとともに、分類のためカテゴリとみなし、ドキュメントを割り付ける方法である。

本論文では、より工学的な問題解決に特化したドキュメントの分類を実現するために、この方法を拡張する。すなわち、変圧器の構造、症状、改修設計行為に関する 3 種類のオントロジーに含まれる語彙をもとに特徴ベクトルを作成し、検討のフローを 3 つの視点から検索可能な形式に分類を行なう。

分類の手順を以下に示す (図??参照)。

step1 検討フロー単位にドキュメントをまとめる。

step2 各検討フロー (ドキュメント) の特徴ベクトルを作成。

step3 構造、対策、症状の各オントロジーの各カテゴリの特徴ベクトルの作成。

step4 構造、対策、症状ごとに検討フローを分類。

3.5.4 特徴ベクトルの生成

特徴ベクトルは、オントロジーの各概念 (語) の重みを要素として表現されたベクトルである。語の重みは、その出現頻度に応じて定義される。特徴ベクトルには、各検討フローに対するものと、オントロジーの各概念、すなわち分類のためのカテゴリに対するものの2種類がある。いずれのベクトルにおいても、その成分はオントロジー中の各概念に対応している。一般に、オントロジーを構成する概念の集合

$$(term_1, term_2, \ldots, term_t)$$

に対して、あるオブジェクト $object_i$ の特徴ベクトル O_i は

$$O_i = (term_{i1}, term_{i2}, \ldots, term_{it})$$

である。ただし、$term_{ij}$ は $object_i$ における $term_j$ の重みを表す。

本論文では、3つのオントロジーから作成した特徴ベクトルの組を合わせたものをドキュメントの特徴ベクトルとして用いる。すなわち、

$$O_i = (O_{Si}, O_{Ci}, O_{Mi})$$

となる。ここで、O_{Si}, O_{Ci} および O_{Mi} は、それぞれオブジェクト $object_i$ の、構造、症状および対策に関するオントロジーから作成された特徴ベクトルである。

検討フローの特徴ベクトルの生成 検討フローの特徴ベクトルは、基本的にはオントロジーの各概念に対応した成分が、それぞれの語の出現頻度を反映したものとなる。さらに、オントロジーを利用することによって、下位概念をもつ概念については、下位概念の出現頻度もベクトル成分に反映させる。すなわち、オントロジーの各概念をキーワードとし、先に作成したページの頻度リストとキーワードとのマッチングを行ない、マッチしたキーワードについて、その出現頻度を対応する成分の値とする。さらに、そのキーワードがオントロジーにおける上位概念を持つ場合は、上位概念に対応する成分にもその出現頻度を加える。こうして

できたベクトルを正規化したものを、その検討フローの特徴ベクトルと定義する。すなわち、ある検討フロー DOC_i の特徴ベクトル D_i は、

$$D_i = \frac{(term_{i1}, term_{i2}, \ldots, term_{it})}{\sqrt{\sum_{j=1}^{t}(term_{ij})^2}}$$

$$term_{ij} = (term_j\text{の出現頻度})$$

$$+(term_j\text{の子概念の出現頻度の総和})$$

で表される。

カテゴリの特徴ベクトルの生成　カテゴリはオントロジー中の概念に対応しているので、カテゴリの特徴ベクトルは概念の特徴ベクトルと言い換えることができる。カテゴリの特徴ベクトルは、対応する概念を表す単語が出現する検討フローの特徴ベクトルを平均したものと定義する。すなわち、オントロジー中のある概念 $term_i$ が出現する検討フローの集合を $(\mathrm{DOC}^i_1, \mathrm{DOC}^i_2, \ldots, \mathrm{DOC}^i_m)$ としたとき、$term_i$ に対応するカテゴリ CTG_i の特徴ベクトル C_i は、

$$C_i = \frac{\sum_{k=1}^{m} {D^i}_k}{m}$$

で表される。ただし、${D^i}_k$ は DOC^i_k の特徴ベクトルである。

分類結果の一例　図??は、「油洩れ」が発生する変圧器の各部品に対するすべての検討フロー45件のうち、「O-リングを取り替える」という対策をとるものを変圧器の部品ごとに分類した結果の上位2件を示している。この結果では、油洩れが生じた場合に、O-リングを取り替えるという対策を取る部品には、ガラス部という類似した構造が存在しすることが分かる。さらに、このような構造の部品から油洩れが生じた場合には、油洩れの程度が「にじみ」でも「油滴」でも、同一の対策を施せばよいことも分かる。このように、オントロジーを用いることで、ドキュメントに明示的に書かれていない専門家の知識を新たに発見することが可能になる。また、変圧器の構造、症状および対策の3つ視点からのドキュメントを検索することができ、内容理解だけでなくメンテナンスの支援にも有効であると考えられる。

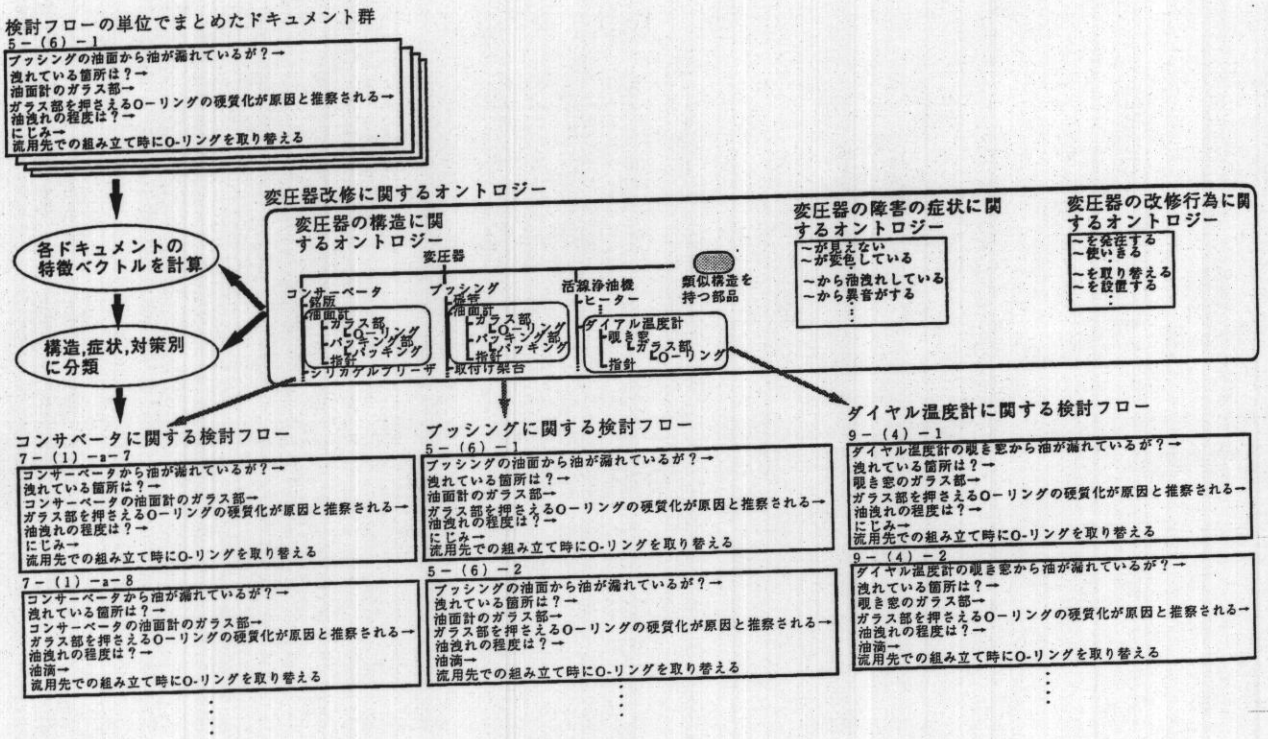


図 3.13 オントロジーによる知的検索

3.6 考察

本研究では現場業務支援システムの構築を現場技術知識の共有化と捉え、オントロジーに基づくドキュメントベースの構築として実現した。ここでの着眼点は、ドキュメントとシステムの統合とオントロジーの利用の2つである。

ドキュメントとシステムの統合には2つの意味がある。まず、マニュアルなどのドキュメントに含まれる知識を積極的に利用してシステムを構築したという点である。システムの基本的動作は、インデントや矢印などの弱い構造をもったドキュメントを変換したものによって作られている。これにより、普通の文章に近い形式でノウハウが記述可能となり、知識記述が簡単になる。さらに記述内容の改訂も容易になる。また、どのような知識が入っているのかも理解しやすくなるので、現場でシステムを使い始めるときの抵抗感を少なくすることができる。2つ目の意味は、システム内の情報はハイパーテキストによって記述されているので、容易に他の情報源と結びつけることができる。これにより、各種規格書などのマニュアルの該当部分といった他のドキュメントや、適宜現場で得られる画像・動画なども連携されることができる。

ドキュメントからのシステム構築ではそのドキュメントの目的に即した知識しか得られないことが多い。これを補うのがオントロジーの利用である。オントロジーを用いることで、対象に関して全般的な知識を付加することができる。すなわち、ドキュメントを対象の性質や構造などによって構造化して提示することができる。このような構造化は、当該の業務遂行に必要な背景的な知識を学習したり、対象の変化や追加に即したドキュメントを修正したりする際に有効になる。

今回作成したプロトタイプは、各変電所、電力所内の約80名の現場技術者にモニターしてもらったところ、「実用において最低限の機能を備えている」、「改修設計時に検討項目の見落し防止になる」などの良い評価を得た。また、変圧器改修業務以外の他部門の事例への水平展開の要望もあった。

一方で、現場での使い勝手の向上について、次のような改善に関する意見が出された。

1. 資料や写真などの表示時に、画面のサイズを大きくする機能を追加する

2. 次にどの画面を見ればよいかをガイドするような機能を追加する

(1) に関してはシステムの実装の問題であり、解決は比較的容易であると思われる。(2) に関しては、現在、Microsoft Agent[?] を利用し、表示画面のナビゲーションを行なう簡単なインターフェースエージェントを実装している。今後はオントロジーと連携したより知的なナビゲーションが課題である。

3.7 3章のまとめ

本研究では、知識獲得や保守のコストの問題から従来のエキスパートシステムでは実現が困難であった、定型・非定型業務が渾然一体となった業務の支援方法を提案した。提案方法では、ドキュメントベースによる「ワープロによる文書作成」の感覚に近い「ドキュメントベース」という簡単な仕組みで知識を記述し、それらをイントラネット上で共有する。これにより個人のノウハウとして蓄積された知識を随時簡単にシステムに組み込み、皆で共有して利用していくことが可能になった。また、ドキュメントベースの構築や保守をより容易にするために、対象業務に関するオントロジーを用いた、技術情報の構造化・共有法を提案した。

また、配電用変電所変圧器設備の改修計画業務支援を対象としたプロトタイプを作成した。本システムはベテラン技術者からの実用的に耐えると評価を受けている。すなわち、本手法は実際的なシステム構築に利用可能であることということが示された。

第4章

オントロジー獲得

4.1 はじめに

2、3章では、知的情報共有を実現するためのオントロジーの利用法に焦点をあて、ネットワークからの情報収集・分類・抽出システム IICA および現場技術共有支援システム OnTheSpot の実装・評価を行い、知的情報共有実現のためにオントロジーが有効であることを示した。

しかし、IICA や OnTheSpot で利用しているオントロジーは、すべて、既存の辞書や専門家が作成したドキュメントなどから手作業で構築しており労力を要する。また、構造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という問題もある。オントロジー獲得支援は、オントロジーの利用において欠くことのできない課題である。

辞書の概念獲得やオントロジー構築に関する研究は、既にいくつか行われているが [?, ?]、その多くは、汎用性を指向するあまり、どのように利用するかという視点に欠ける。このため各研究で提案されている手法に基づいて獲得された辞書やオントロジーがどれほど有効なものであるか客観的に評価を行っている例は少ない。

本章は以下のように構成されている。そこで、本章では、利用面を考慮入れたオントロジーの獲得支援法について、(1) 概念間の関係の獲得, (2) 概念の獲得の2つの問題に分けて検討する。

(1) に関しては、収集した WWW ページから、特徴ベクトルの類似度、用語の共起関係、リンク情報を利用した3種類のオントロジー獲得法を用いて、旅行、情報科学の二つの分野におけるオントロジー獲得の実験を行ない、得られた結果

をもとにオントロジー獲得の可能性について考察する。

(2) に関しては, 分類された WWW ページから抜き出した単語の情報量からその単語の重要度を決定し、概念の獲得方法を提案する。提案法によって得られた概念に基づく、特徴ベクトルを利用した分類実験を行ない、求められた構成要素の有効性を評価する。関連のページ分類についても本方法を適用し、(1) の結果との比較検討も行なう。

4.2 概念関係の獲得

ここでは、WWW ページから得られる情報を利用した 3 種類の概念関係の獲得支援方法について検討する。

4.2.1 類似度からの獲得

2 章で述べた WWW ページの分類の過程で得られる各概念の特徴ベクトルを用いて、ページと概念との類似度の計算と同様に、概念間の類似度を計算することができる。類似度の高い概念の間には、何らかの意味的な関係が存在すると考えられる。したがって、ある閾値を超える類似度をもつ二つの概念に関係を与えていくことにより、対象領域に関する新たな概念間の関係を得ることができる。ただし、ここでは概念間の関係は与えられていないことを前提としているので、WWW ページ分類の処理とは異なり、ページの特徴ベクトルは概念そのもののマッチングのみを利用して作成するものとする。

アルゴリズム

1. オントロジーの構成要素となる概念の集合を用意する。
2. 各 WWW ページについて、 2.5.2 項と同様の方法で特徴ベクトルを生成する。ただし、$term_{ij}$ = (ページ $page_i$ における概念 $term_j$ の出現頻度) である。
3. 各概念について、2.5.2 項と同様の方法で特徴ベクトルを生成する。

4. すべての二つの概念の組合わせについて、概念間の類似度を計算する。類似度は二つの概念の特徴ベクトルの内積で定義される。

5. 類似度が設定した閾値を超える概念の組を抽出し、その集合を出力する。

4.2.2 用語の共起関係からの獲得

この手法は、概念の特徴ベクトルを生成する部分までは類似度からの獲得と同様であるが、各概念について、その特徴ベクトルの成分のうち設定した閾値を超えるものに対応する概念と関係づける、というものである。これは、一つのWWWページに共に出現する用語の間には何らかの関係がある、という考えに基づくものである。

アルゴリズム

1. 類似度からの獲得の場合と同様に、概念の特徴ベクトルを生成する。

2. 概念 i の特徴ベクトル $(term_{i1}, term_{i2}, \cdots, term_{in})$ に対する閾値を

$$\frac{1}{\sum_{k=1}^{n} term_{ik}}$$

と定義する。ただし、n は用意された概念の数である。この閾値は $\frac{1}{\sqrt{n}}$ から1までの値をとり、ベクトルの成分の分散が大きいほど閾値は大きくなる。

3. 各概念について、その特徴ベクトルの成分のうち2で計算された閾値を超えるものを検出し、その成分に対応する概念を、もとの概念と関係づける。ただし、もとの概念自身に対応する成分を除く。

4.2.3 リンク情報からの獲得

WWWページの場合、ページ間のハイパーリンクと概念間の関係について、何らかの相関があるのではないかと考えられる。そこで、収集・分類した旅行関連のWWWページのうちで、互いにリンクで結ばれている関係にあるものを調べ出し、さらにリンクで結ばれた二つのページが、それぞれどのカテゴリに分類さ

れたかを調べる。そして、それらのページが属するカテゴリが異なる場合には、その二つのカテゴリには何らかの関連があるとみなすものとする。全てのリンク情報に対してこのような操作を施すことにより、新たな概念間の関係を得ることができる。ただし、既出の二つの手法と同様オントロジーは与えられていないことを前提としているので、分類はキーワードマッチングのみを利用して作成した特徴ベクトルを用いて行なう。

アルゴリズム　ページ間のリンク情報からのオントロジー獲得のアルゴリズムを以下に示す。

1. ページを収集する段階において、各ページの URL と、そこからリンクを張ってあるページの URL（のリスト）を、ページの ID に対応させてそれぞれ記憶する。相対アドレスで書かれてある場合は、ページ自身の URL をもとに絶対アドレスに変換する。
2. 1 ページずつ順に、そのページからリンクを張っているページが収集したページの中に存在しているかどうか調べ、あればリンク先のページの ID を保持していくことによって、リンク先のページ ID のリストを作る。
3. 3 章と同様の方法によって、収集したページの分類を行なう。ただし、ページの特徴ベクトルの計算において、$term_{ij}$ = (ページ $page_i$ における概念 $term_j$ の出現頻度) である。
4. リンクで結ばれた二つのページがそれぞれ属している概念が異なる場合には、それらを関係づける。
5. 4 をすべてのリンクに対して実行することにより、新たな概念の体系を構築する。

4.2.4　オントロジー獲得実験

実験 1 - 旅行関連データへの適用　前節で紹介した 3 種類の方法を用いて、WWW ページからの概念関係獲得実験を行なった。実験データには、??項の分類実験

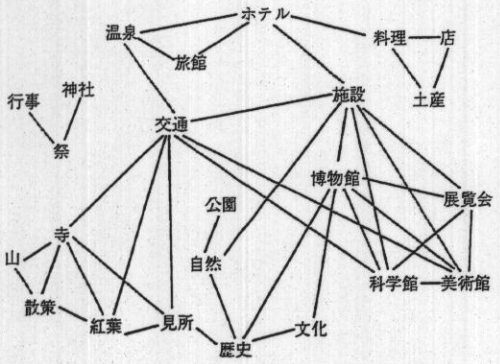


図 4.1 類似度から得られた旅行のオントロジー

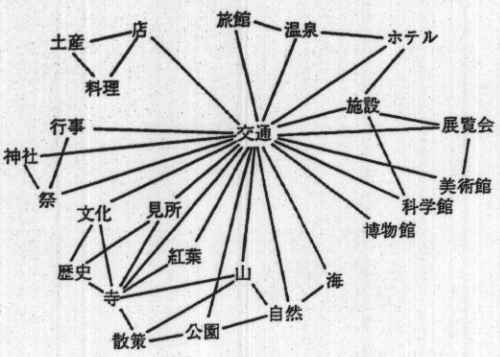


図 4.2 用語の共起関係から得られた旅行のオントロジー

において使用した旅行に関する WWW ページ使用した。類似度、共起関係およびリンク情報から得られた旅行に関するオントロジーを図??、図??および図??に示す。

評価　また、獲得されたオントロジーの妥当性を評価するために、3 章で手作業で作成した旅行のオントロジー (図??参照) と比較し、得られた概念間の関係が適切なものであるかどうかを調べた。評価基準としては、以下の 3 種類 (○、△、×) を用意した。

○：実際に関連がある

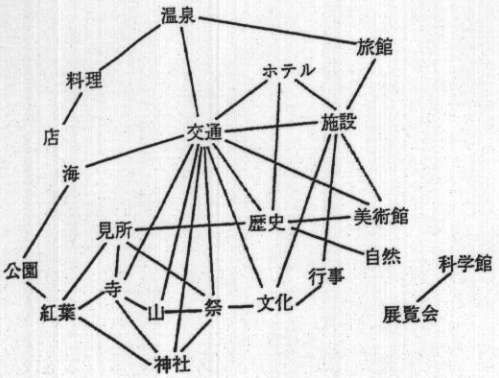


図 4.3 リンク情報から得られた旅行のオントロジー

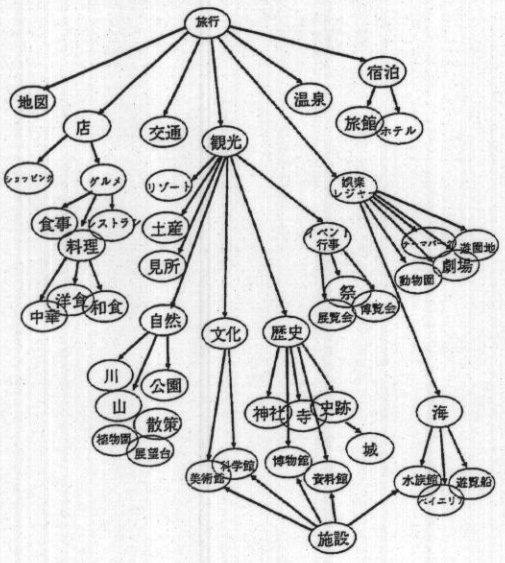


図 4.4 旅行のオントロジー

表 4.1 用語の共起関係から得られた旅行のオントロジーに対する主観評価

評価	○	△	×
類似度から得られた該当する関係の数	76	29	13
共起関係から得られた該当する関係の数	90	19	14
リンク情報から得られた該当する関係の数	35	33	12

△：直接的な関連はないが、無関係ではない

×：無関係である

得られた概念間の関係に対する評価を表??に示す。この評価結果に関する詳しい議論は、第 4.2.5 項の考察で行なう。

実験 2 - 情報科学関連データへの適用 同様の実験を、情報科学関連の WWW ページを対象にして行なった。実験に用いたのは、大学の研究室の研究内容を紹介している WWW ページ約 100 件である。

評価 獲得されたオントロジーとの比較評価を行なうために、専門文書に基づいて手作業で作成されたオントロジーを用意する (図??参照)。このオントロジーは、岩波情報科学辞典を参考にして作成したものである。ここでは、獲得されたオントロジーとの比較に用いるため、構成概念は必要最小限にとどめてある。

類似度および共起関係から得られたオントロジーと情報科学辞典に基づいて作成したオントロジーとの違いをそれぞれ図??、図??に示す。ただし、図??における細線は類似度から得られた概念間の関係を表し、太線は情報科学辞典に基づいて作成したオントロジーにおける概念間の関係を表す。また、図??において、細線は共起関係から得られた概念間の関係を表し、太線は情報科学辞典に基づいて作成したオントロジーにおける概念間の関係を表している。

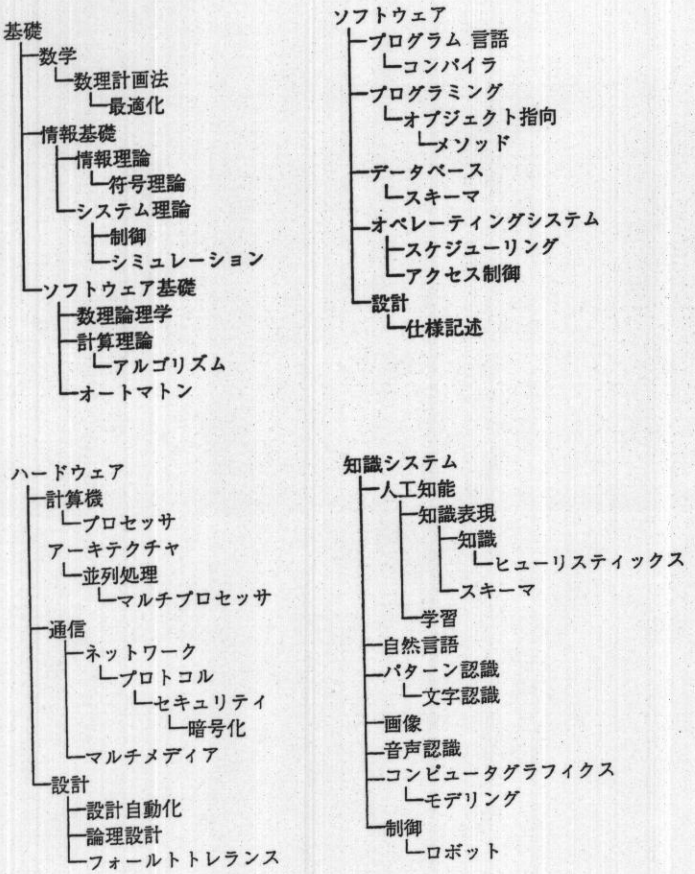


図 4.5 岩波情報科学辞典をもとに作成した情報科学のオントロジー

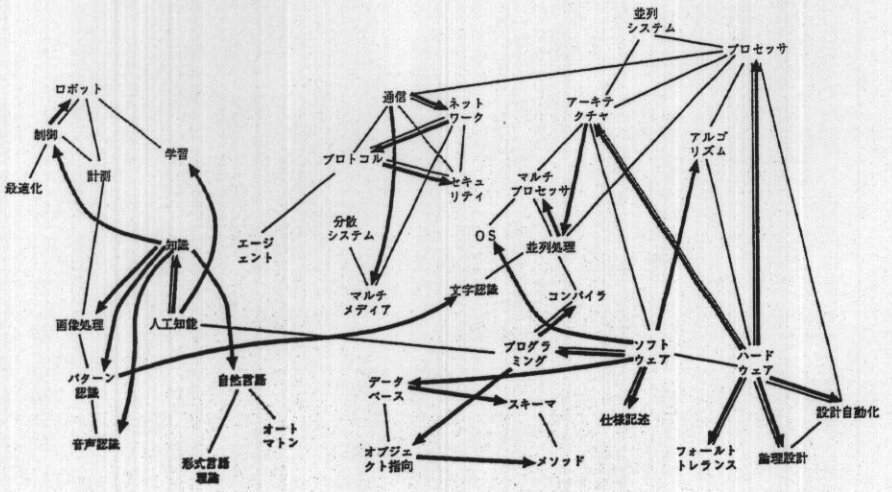


図 4.6 情報科学辞典から作成したオントロジーと類似度から得られたオントロジーとの比較

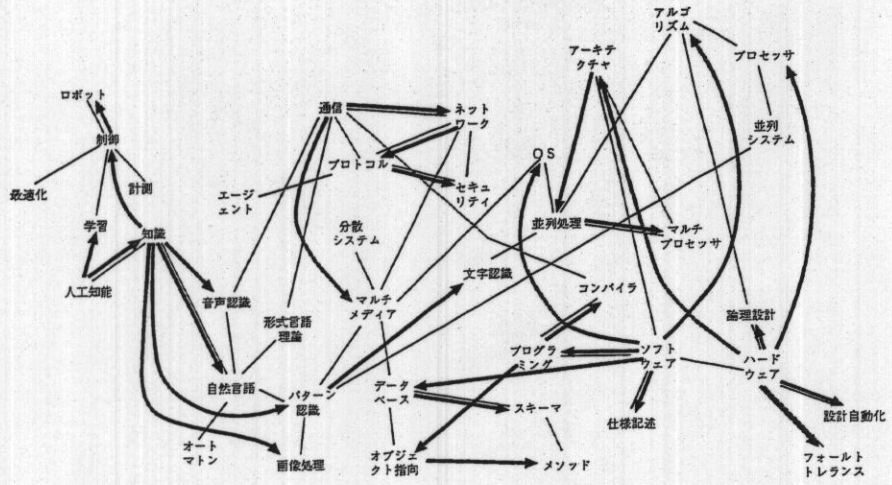


図 4.7 情報科学辞典から作成したオントロジーと共起関係から得られたオントロジーとの比較

4.2.5 考察

ここでは前節で行なった概念関係獲得実験の結果に基づいて、オントロジー獲得に関する考察を行なう。

各手法の比較 まず、類似度からの獲得、共起関係からの獲得について考える。これらの手法は、ともに同じ特徴ベクトルをデータとして処理が行なわれている。

表??より、類似度からの獲得、共起関係からの獲得とも、得られた概念間の関係のうちの多くが適切なものであったことがわかる。ただし、ここでは本来与えられるべき関係のうち実際に与えられた関係がどの程度であったか、という評価をしていない。これは本来与えられるべき関係というものを定義することが困難であるためである。したがって、これらの手法から得られたオントロジーが必要な概念間の関係のうちのどの程度を網羅しているのかを評価することはできないが、得られた関係の精度の高さにおいて、これらの手法は有効であると言える。

類似度からの獲得では、局所的に完全グラフに近いような構造が現れる。これは、互いに類似している概念のグループがいくつかでき、それらのグループ内では互いの類似度は高くなっていることによるものである。この場合、他の経路によって結ばれている概念間にも、冗長な直接のリンクがつくことになる。また、グループ間のリンクができずに孤立してしまう部分が多く、オントロジーが一つの体系にならない。一方、共起関係からの獲得では、概念間の連想関係が忠実に反映されており、一部の概念だけで閉じてしまうことなく、全体的な概念間のつながりを表現することができる。

また、類似度からのオントロジー獲得の場合、一定の閾値を設定しなければならない。ここでは、予めすべての二つの概念間の類似度を調べることにより、0.75という値を採用している。しかし、扱っている対象が人間の書いた文書という不確定要素の強いものであり、一定の閾値をあらゆる分野のあらゆる文書に対する類似性の境界とするのは問題である。分類結果などから閾値を学習し、動的に変化させる方法が考えられるが、具体的な学習法は今後の検討課題である。これに対して共起関係からの獲得の場合、閾値を定式化することができ、実験の結果この閾値が妥当なものであることが判明した。

次に、リンク情報からの獲得について考える。リンク情報からの獲得では、多くのWWWページ作成者の対象領域に関する知識を反映させることができると考えられる。得られた概念間の関係に対する評価は表??に示すようにさほど良くない。しかし、図??～図??を比較すると、例えば、温泉と料理との関係性については、リンク情報からのみ得られおり、類似度からの獲得、共起関係からの獲得と違う側面の関係を見い出すが可能であることを示唆している。

概念間の関係　ここでは、獲得された概念間の関係について考察する。オントロジー獲得のためのそれぞれの手法における概念間の関係について、次のような問題点が挙げられる。

類似度からの獲得の場合

関係づけられた二つの概念は対等なので、これだけでは上位・下位関係などを定義することはできない。

共起関係からの獲得の場合

関係づけられた二つの概念には特徴ベクトルとその成分という主従関係があるが、これを直接上位下位関係と定義することが妥当であるとは、一概にはいえない。また、この主従関係のために、双方向の関係を生じる場合がある、という問題点もある。やはりこの方法でも、機械的に上位・下位関係を定義するのは難しいと思われる。

リンク情報からの獲得の場合

親ページ、子ページという明確な主従関係があるが、これも共起関係と同様双方向の関係を生じる場合がある。Yahoo![?] などのインデックページのみを集めたデータから統計をとることができれば、概念同士の上位・下位関係を推定することはある程度可能であると思われる。

以上のように、現状では得られた概念の関係がどのようなものであるかということを機械的に定義することは困難であり、単に連想関係として扱わざるを得ない。しかし、WWWページのリンク情報、共起関係、類似度の3つ情報を用い

ることで、それぞれ違った側面の関係を見い出すことが可能であることがわかった。詳細な関係の意味を定義する必要がある場合は、獲得されたオントロジーに手作業で意味づけを行なわなければならないが、関係そのものは既に与えられており、本手法を用いることによってオントロジー作成における負担は軽減されると思われる。

用途に合ったオントロジーの使い分け　学問のような、意味がある程度安定した分野を対象とする場合、既に専門家によってトップダウン的に作られたオントロジーが存在することもあるので、それを利用すればよい。ただし、対象領域において新たな概念が生じた場合に、本章で獲得されたオントロジーが利用できると考えられる。例えば図??において、既存のオントロジーには存在しなかった「エージェント」が、獲得されたオントロジーによって「プロトコル」の関連語であるということがわかり、この知識を既存のオントロジーに組み込むことが可能である。

旅行のような、用いられる語の意味が安定していない分野を対象とする場合、主観的な判断によって概念間の関係を一から構築していくことは効率的ではない。このような場合、統計的・機械的に作成されたオントロジーを利用することは有用である。

4.3　概念の獲得

本節では、オントロジーの概念候補を抽出する手法として、対象となるデータ (WWW ページ) から抽出した単語の情報量を用いて、その重要度を決定する方法を提案する。また、この方法の有効性を評価するために、決定した重要単語に基づいて作成した特徴ベクトルによる WWW ページの分類実験を行なった。

4.3.1　情報量に基づく単語の重要度 (情報価値) の決定方法

本論文では、単語の重要度の基準 を**情報価値**と呼ぶことにする。

WWW ページから、オントロジーの獲得支援を行なうために、WWW ページを解析する必要がある。WWW ページからの重要単語の決定方法は、Syskill & Webert[?] の考え方に基づいている。

情報価値の決定手順は以下の通りである。

step1　対象に関連するWWWページを収集。

step2　各WWWページについて形態素解析を行ない、名詞のみを表??のルールに従って文字や単語の連結を行ない、単語リストを生成。

step3　各ページがある概念に関連するか否かを決定(手動)。

step4　各WWWページの単語リストから、各単語の情報量を求めることで単語の重要度を決定。

step5　step3, step4を各概念毎に繰り返す。

エントロピーの計算　いま注目しているの一つ単語（概念）に、WWWページが関連しているときをhot, 無関係なときをcoldで表し、その単語がWWWページに出現しているときをpresent, 出現しないときをabsentで表すとすると、エントロピーの計算方法は、

$$E(W,S) = -P(W = present)I(S_{w=present}) - P(W = absent)I(S_{w=absent})$$

ただし、

$$I(S) = \sum_{C \in \{hot,cold\}} -p(S_c)log_2(p(S_c))$$

である。

図??は、ある単語が目的のWWWページにどのような分布で出現するのかを摸式的に表したものである。この例では、全部で100個のWWWページがあるとしてる。いま左の図は、注目している単語が100個のWWWページの中に20回出現し、そのうち目的のページに出現する回数が10回、目的のページでないページに出現する回数が10回である場合である。このとき、エントロピー=0.85と計算される。一方、図??の右側の場合は、同じ回数出現したが分布の異なる単語の場合を示している。目的のページに出現する回数が15回、目的のページでないページに出現する回数が5回であり、エントロピー=0.71となる。したがっ

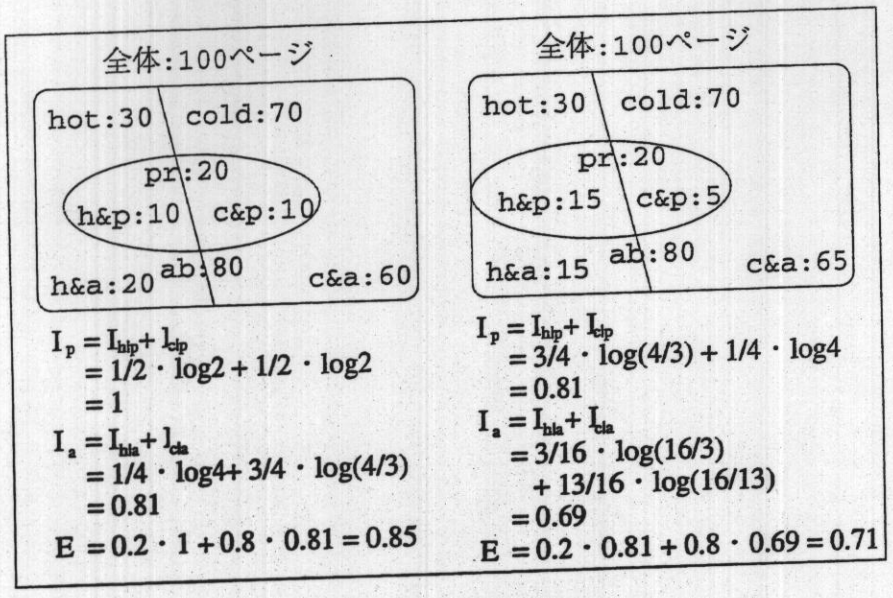


図 4.8 キーワードの選択

て、この例では、右側の単語の方が情報量という視点から重要であることがわかる。つまり、エントロピーが小さい方が、その単語が偏って出現することを表し、情報価値が高いということになる。

情報量が小さくなる場合には、図??のように、*hot* なページに偏って出現する場合と、その逆で、*cold* なページに偏って出現する場合が考えられる。本研究で情報価値が多い単語という場合には、後者の場合を考慮に入れていない。その理由は、*cold* に分類されるページには他の概念の *hot* なページが収集されているため、後者は起こらないと考えたからである。

4.3.2 概念獲得実験

上で述べた方法に基づいて、概念獲得実験を行なった。表??は、実験に使用したデータ数 (WWW のページ数) を各概念ごとに示している。

実験によって得られた概念が、適切なものであるか調べるために、縦軸をユーザがカテゴリに属すると判別した WWW ページの数 (Y)、横軸を情報価値の高い上位 n (= 10, 20, 30) 個を多く含んでいる順に WWW ページをソートしたと

表 4.2 各概念に分類されたページ数

カテゴリ	ページ数	定義
CD-ROM	33	CD-ROM
PRINTER	53	プリンタ
SCANNER	30	スキャナ
WORD-PRO	21	ワードプロセッサ
MO	32	光磁気ディスクドライブ
MODEM	46	モデム
TA	41	ターミナルアダプタ

きの WWW ページの数 (X) として、プロットした。その結果の一部を、図??〜図?? に示す。

このグラフでは、Y = X となったとき、システムが選ぶページがユーザが選択したページが完全に一致したときであり、最も理想な場合となる。各グラフから明らかなように、情報価値の高い単語が 10 個の場合と、20 個、30 個の場合では、グラフの傾きに開きが見られた。また、グラフには示していないが、情報量の高い単語を 30 個以上使用しても大きな変化見られなかった。この方法で獲得した概念をオントロジーを利用して WWW ページを分類する場合は、情報価値の高い単語を上位 20 個〜 30 個程度使用する場合が一番効率が良いといえる。図??は、図??のデータを用いて、横軸を適合率、縦軸を再現率に、プロットしたものである。

4.3.3 評価実験 (獲得概念による WWW ページの分類実験)

本項では、情報量により概念獲得法の有効性の評価を、利用面も考慮に入れて行なう。

評価は、情報量によって選出した単語から各カテゴリの特徴ベクトルを生成し、WWW ページの分類実験を行ない、その適合率、再現率を求めることによって行なう。これは、情報量によって選ばれた単語に基づいて生成された特徴ベクト

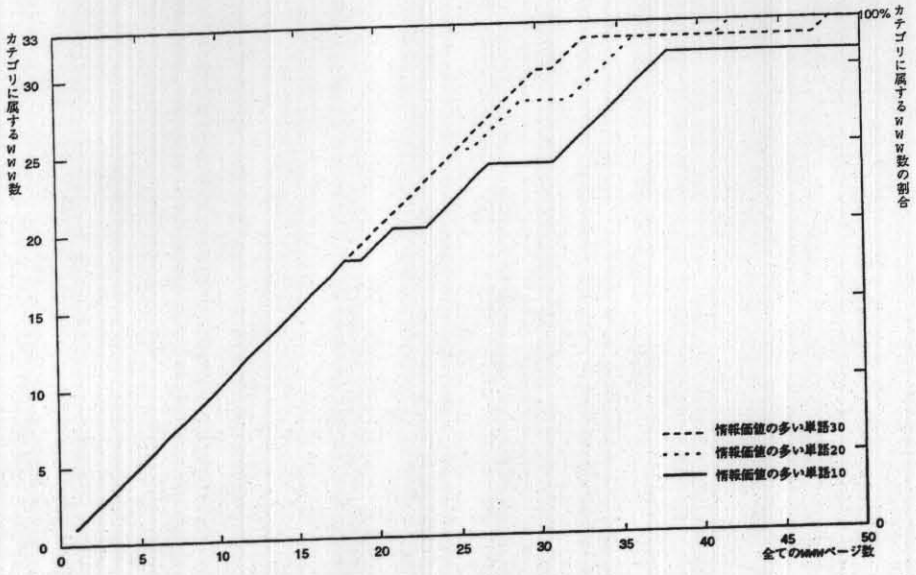


図 4.9 構成要素の包含数 (CD-ROM)

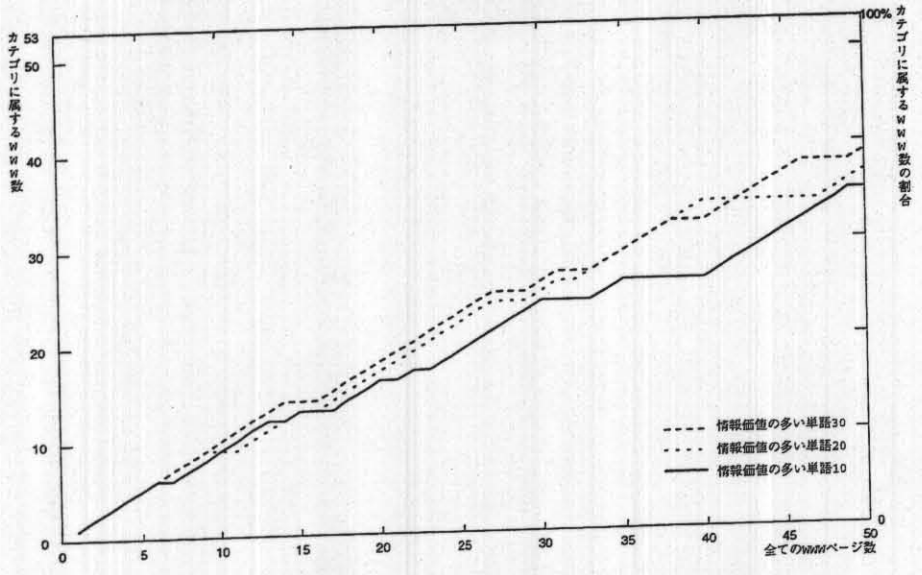


図 4.10 構成要素の包含数 (PRINTER)

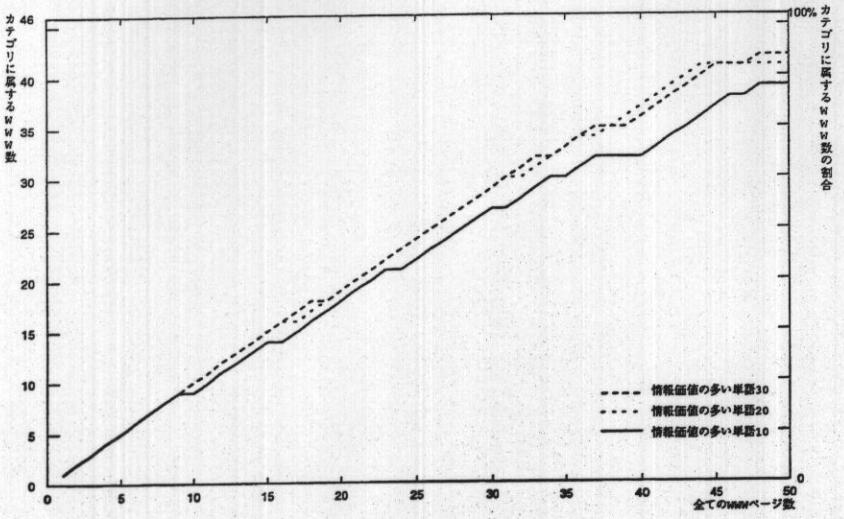


図 4.11 構成要素の包含数 (MODEM)

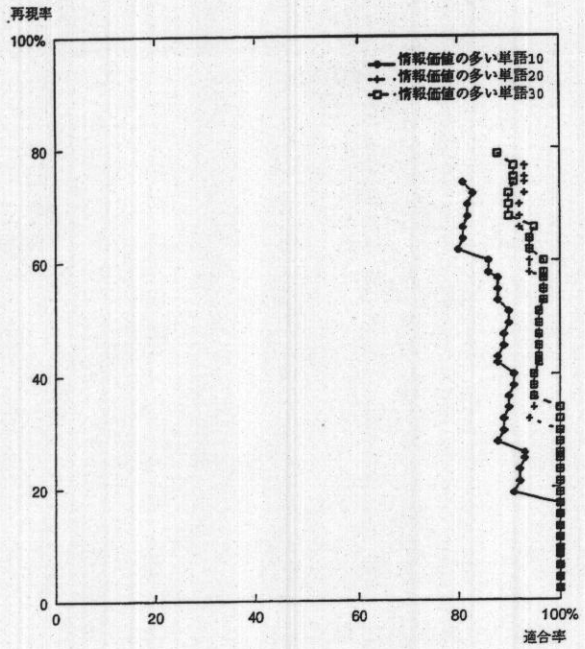


図 4.12 構成要素の包含数による適合率と再現率 (MODEM)

ルによって、効率よく分類が可能であれば、それらの単語は妥当性の高いものであり、提案法が有効であることが示せると考えられるからである。ここで、特徴ベクトルの生成および WWW ページ分類は、2.5 節で説明した方法にしたがっている。

実験は、前項で選出した情報機器関連の単語を用いて行なう。また、2.5 節のオントロジーによる分類実験で用いた旅行データにも適用し、手作業で選んだ単語に基づく分類結果と情報量によって選んだ単語に基づく分類結果との比較検討を行なう。

実験 1 - 情報機器関連データへの適用　分類実験は次の 2 種類のデータに対して行なった。

実験 1-A 情報量の計算による単語を選出のために使用した WWW ページ (教師データ)

実験 1-B 情報量の計算に使用していない、別の WWW ページ (テストデータ)

ここで、分類のために作成した特徴ベクトルの要素数は、137 個（7x20 = 140、そこから重複を省く）である。

実験 1-A および実験 1-B の評価結果を表??に示す。テストデータに対する分類においても教師データと同等の適合率、再現率が得られた。このことから、WWW ページから抽出された情報価値の高い単語が WWW ページを分類に対して有効であることがわかった。

この実験により、各概念ごとに多少ばらつきがあるがかなり高い確率で分類できていることが分かる。適合率、再現率は 2.5.5 項で定義したものと同じである。ページ数とは、各カテゴリの基になるページ数である。

実験 2 - 旅行に関するデータへの適用　本実験では、2.5.5 項の分類実験で用いた旅行に関するデータ (WWW ページ) を使用し、2.5.5 項の分類精度とのを比較行なう。

本実験は、実験 1 同様、次の 2 種類のデータに対して分類を行なった。

表 4.3 WWW ページ分類実験の評価結果

カテゴリ名	実験 1-A		実験 1-B	
	適合率	再現率	適合率	再現率
CD-ROM	89 %	100 %	92 %	94 %
PRINTER	96 %	92 %	91 %	78 %
SCANNER	100 %	90 %	82 %	96 %
MO	85 %	94 %	66 %	86 %
MODEM	93 %	93 %	92 %	66 %

実験 2-A 情報量の計算による単語を選出のために使用された WWW ページ (教師データ)

実験 2-B 情報量の計算に使用していない、別の WWW ページ (テストデータ)

図??は、2.5.5 項の実験結果、実験 2-A および実験 2-B の評価結果を、縦軸に適合率、横軸に再現率として、カテゴリごとにプロットし、各実験毎に線で囲んだものである。この図より、情報量に基づいて選出した単語から作成した特徴ベクトルによる分類精度が、手作業で作成した特徴ベクトルによる分類精度と同等であることが判る。

4.4 考察

現在のところ、概念間の関係が明確に定義できない、無関係な概念どうしが関係づけられる場合がある、などの問題が残っており、この点で獲得されたオントロジーをそのまま利用することは難しい。しかしながら、問題のある部分を手作業で修正することは、オントロジーを一から手作業で作成することに比べればはるかに負担が少なく、現段階でもオントロジー作成支援ツールとしての役割は十分果たし得る。

共起関係からの獲得とリンク情報からの獲得は全く異なるアプローチによるものであり、この二つの手法を利用して、より実用的なオントロジーを獲得するた

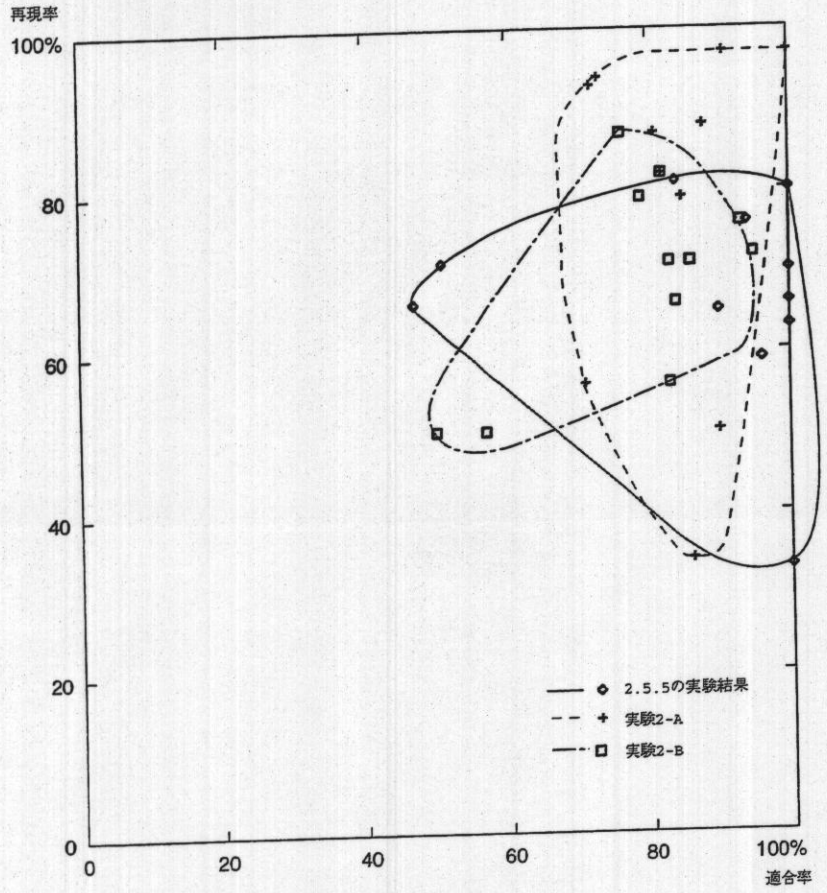


図 4.13 カテゴリの適合率と再現率の分布

めの枠組を実現することが期待される。

またこれらの手法により、新たにWWWページを収集した場合、それらのページを既存のデータに含めて特徴ベクトルの再計算、ハイパーリンクの検出を行ない、オントロジー獲得の処理を再実行することによって、新しい情報の内容を反映したオントロジーの変更が可能である。

4.5 第4章のまとめ

本章では、利用面を考慮入れたオントロジーの獲得支援法について、(1) 概念間の関係の獲得, (2) 概念の獲得の2つの問題に分けて検討した。

(1) に関しては、収集したWWWページから、特徴ベクトルの類似度、用語の共起関係、リンク情報を利用した3種類のオントロジー獲得法を用いて、旅行、情報科学の二つの分野におけるオントロジー獲得の実験を行ない、得られた結果をもとにオントロジー獲得の可能性について考察した。

(2) に関しては, 分類されたWWWページから抜き出した単語の情報量からその単語の重要度を決定し、概念の獲得行なう方法を提案する。提案法によって得られた概念に基づく、特徴ベクトルを利用した分類実験を行ない、求められた構成要素の有効性を評価する。旅行関連のページ分類についても本方法を適用し、(1) の結果との比較検討を行なった。

第5章

関連研究と考察

5.1 はじめに

本章では、オントロジー、インターネットロボット、エージェント, テキスト分類、内容処理および工学的知識の共有という広範囲にわたる観点から関連研究と比較し、本研究の特色について述べる。また、オントロジーに基づく情報共有のアプローチの特色についていくつかの疑問点に答えていく形式で総括的な議論を行なう。

5.2 関連研究

本節では、本研究と関連する従来の研究との比較を行ない、本研究の特色について述べる。

5.2.1 オントロジー

オントロジーに関する研究には、CYC[?] や Sharable Ontology Library[?] 等がある。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。

本研究で用いたオントロジーはドメイン・オントロジーとタスク・オントロジーとみることができる [?]。オントロジーに関する研究は多数あるが、ここでのオントロジーは1章で述べたように各種表現の統合のための利用という点で特徴的である。今後はより体系的なオントロジー構築について考慮していく必要がある。

こうした研究に、タスク・オントロジーについては高岡ら [?] の研究において具体的な 事例に基づいてオントロジーの構成方法について議論を行なっている。

一方、オントロジー獲得に関しては、松尾らの科学技術論文からのオントロジーの自動生成方法 [?] は、オントロジーを語彙の順序集合と考え、有限順序集合で表現し、語彙の類似性を位相数学の開集合によって規定した場合の数学的性質からオントロジーの形式化を行なっている。この手続きは、基本的に開集合の包含関係に基づいている。しかし、類似した概念の集まりを開集合と仮定し、すべての開集合が観測されたと仮定しているため実用に耐えられる段階ではない。本研究におけるオントロジー獲得法は、同じく本研究で提案しているオントロジーによる分類に利用可能であり、実用を十分考慮している点に特徴がある。

5.2.2 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 [?] 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 [?, ?] 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。

5.2.3 テキスト分類

シソーラスのような構造化知識を利用した、テキストの自動分類の研究はすでに多く行なわれている。山本らは、分類体系相互の関係を利用した日本語テキストの自動分類手法を提案した [?]。河合は、単語表記の統計的な情報を用いた方法を、分類体系と独立に作成した意味属性体系に適用するテキスト自動分類方式を提案した [?]。また、亀田らによる、テーマ・キー概念・キーワード間の階層構造を利用する新聞記事情報の分類・検索システムに関する研究もある [?]。

しかし、これらのアプローチでは, シソーラスが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。

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[?, ?]。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もとづいてカテゴリーベクトルおよびオントロジーの概念間の重み付けも変更より、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。

5.2.4 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[?]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[?]の報告などが挙げられる。

電子ニュースのダイジェスト生成システム[?]では、ニュースグループfj.meetingsの会議告知記事や論文募集記事からタイトル、開催期日、開催場所、論文締切、などを抽出する。抽出する方法は、文章のスタイル情報と言語パターンを使用し、対象項目がどの位置に記述されているかを解析し、抽出を行なう。

自動抄録生成システム[?]では、文書構造(章や節などの書式)と修辞構造(文と文の修辞関係)を解析し、技術論文、社説を対象にした抄録の自動生成を行なう。このシステムでは、修辞構造を利用して、文に重みをつけていき、一番低い文を重要文として文章から抜きだし再構成して提示するシステムである。

松尾ら[?]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。

また、米国の DARPA の研究プロジェクトの一環として MUC (Message Understanding Conference)[?] が研究のコンペティションを実施してきた。これは共通のテキストを材料として、各研究で作成した情報抽出システムの能力を競うものである。このコンペティションに参加したシステムが使っている処理方式は簡単な構文解析、意味解析を経た後に言語表現パターンとのパターンマッチングを行なうものが主であった [?]。

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・統合化を行なっている点で、本研究と類似している。しかし、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だけでなく、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、多様な情報源に容易に適用可能である点が特徴である。

5.2.5 工学的知識の共有

電力分野における知識の共有化の試みとしては、田中らが、電力系統の技術者教育を目的に、「共同作業を通して知識の生成、蓄積、伝承を支援する」対話型教育支援システムを提案している [?]。このシステムでは、個人のための教育支援環境に加えてコンピュータを介して離れた人とコンピュータ画面やデータを共有することが可能であり、他者と共同で系統現象を解析したり議論したりすることができる。また検討した事柄や習得した知識を構造化して管理する機能を提供することで、他社や次世代への知識の伝承支援を試みている。しかし、このシステムでは、系統モデルの作成と数値シュミレーションといったモデル的な限定された知識しか扱っていない。

また、高橋らは、ノイズ対策の熟練技術者によって行われたノイズ対策事例を分類し、具体的なノイズ対策手順を知識ベース化することで、ノイズガイダンスシステムおよびノイズエキスパートシステムを構築している [?]。しかし、このシステムはプロトタイプであり、実際的な解決方法を得るためにはより強力な推論プロセスが必要である。また、 知識工学の専門家以外の人が自由に知識を追加・変更できないという問題もある。

より簡単に知識を扱う方法に関する例としては大嶽らの研究 [?] がある。この

研究で提案されているシステムはＯＡ全般を対象とした知識情報共有システムで、ユーザをオフィスワーカ全体としている。一方、著者らの提案システムでは、技術分野に範囲を絞り、ユーザはその分野の専門家（たとえ非熟練者であっても）である。また、対象範囲を限定することにより、知識情報共有に加え、業務支援機能を備えたシステムを指向している。また、電力の運転・保守・補修といった分野では、ノウハウ的でボトムアップ的な知識（個人的で自由度の高い知識）だけでは業務遂行は難しく、トップダウン的な知識（組織化されよく管理された知識）も必要となる点を鑑み、両者の知識の融合を目指した。いわば、芯のある柔らかいシステムを目標としている。

5.3 議論

本節では、オントロジーに基づく情報共有のアプローチの特色についていくつかの疑問点に答えていく形式で議論を行なう。

5.3.1 本研究のオントロジーにおける構造的、意味的不完全性について

一階述語論理に基づいたオントロジーに関する研究が行なわれており、形式的な操作性に関しては優れていることが知られている。しかし、実世界の情報は論理的、意味的に曖昧で不完全なものが非常に多く、これらを述語論理を用いてトップダウン的に構築する方法は、現実的ではない。そこで、本研究では、形式的な操作性は失われるが、より現実の情報に対応するために、概念とその関係および表現との対応程度を持つオントロジーを採用した。本研究におけるオントロジーの主な役割は、表現の統合であり、振舞いの予測などの機能は必要としていないため、概念とその関係のリンクしか持たない弱い構造のオントロジーでも十分その役割を果たしている。

5.3.2 シソーラスと本研究のオントロジーの相違点

シソーラスとオントロジーの相違点は、前者が自然言語と密接に関係しているのに対し、後者は概念と密接に関係にあることである。すなわち、シソーラスは、

言語レベルで語彙を整理してあるのに対し、オントロジーでは、対象領域に関する概念の切り出しとその意味の確定することがその本質となっている。

溝口 [?] は、オントロジーの形態（機能）を

(1) 共通語彙

(2) 概念 (用語) の階層的記述

(3) データベースの概念スキーマ

(4) シソーラス

(5) シソーラス＋目的から見て必要な概念と概念間の関係との厳密な記述

にわけ、(5) をオントロジーの本質であり、(1) ～ (4) を包含するものであるとし、オントロジーを利用法を 3 つのレベルに分類している。

オントロジーの基本的な機能は、対象世界に存在する概念の切り出しとそれらの関係の記述である。そして、最も一般的で簡単な記述が階層関係の記述であり、そこには概念のラベルと階層記述だけがある。これは、最もプリティブなオントロジーであり、第 2 章の IICA におけるオントロジーもこれにあたる。

すなわち、第 2 章で利用したオントロジーや、第 4 章で獲得したオントロジーは、シソーラスと全く異なる概念体系を提案するものではなく、むしろ、シソーラスや専門用語辞書のような既存の概念体系をオントロジーとしてどのように利用していくかということに焦点をあてたものである。第 2 章のオントロジーの構築にあたっては、まず既存のシソーラスをオントロジーの原型としてシステムに適用し、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ていく、という方法をとっている。また、第 4 章で獲得したオントロジーは、第 2 章の IICA システムの利用を目的に、実データから構築したものである。この意味では、我々のアプローチはシソーラスを包含しているといえる。

一方、第 3 章の OnTheSpot における変圧器の構造に関するオントロジーでは、部品の接続や包含といった概念関係を定義しており、単なるシソーラスにとどまらず、オントロジーの本質により近いものであると考えられる。

また、シソーラスが、汎用性を目指して作成されたものであるのに対し、本研究のオントロジーは、対象領域に関する知識の構造化と共有を目的としており、必ずしも汎用性を目指したものではない点が異なっている。

5.3.3 現場技術情報共有支援システムとエキスパートシステムとの相違点

第4章の研究上の興味は作成したシステムそのものにあるのではなく、システム構築の方法論にある。本研究では現場業務支援システムの構築を現場技術知識の共有化と捉え，オントロジーに基づくドキュメントベースの構築として実現している。ここでの着眼点は，ドキュメントとシステムの統合とオントロジーの利用の2つにある。すなわち、ドキュメントを元にシステムを構成するために、普通に文章を書くに近い感覚でシステムに知識を入力することが可能になる。また、対象そのものに関わる知識などはオントロジーとして与えることにより、ドキュメントに表面上現れない知識を補完することができる。このようなシステム構築方法論は新たなアプローチであると考えられる。

技術者が持っている経験的知識を再利用可能な形で記述・整理する必要があるという点では、従来の方法も今回提案した方法も方法を問わず変わらないと思われる。しかし、ここで問題としているのは、それをどのように実現するかという点である。

これまでのエキスパートシステムの構築にあたっては、知識技術者あるいはシステムに習熟したものが細部に渡ってそのシステム特有の知識表現に合わせた記述を生成する必要があった。この部分が、経験的知識を再利用可能な形で記述・整理する段階にあたる。この段階が多くの労力を必要とするだけでなく、システム構築後のシステムの理解のしやすさや更新のしやすさという点で問題があった。

これに対して、第4章の方法では、ドキュメント作成とオントロジーの構築というところが、経験的知識を再利用可能な形で記述・整理する段階にあたる。専門家は業務特有の知識は少数の制約 (インデントや矢印などによるフローの明示) に基づく文章として記述する。オーサリング・ツールはそれをシステムが利用可能な形のハイパーテキスト形式に変換し、システムに渡すことで、システムへ知識の入力がなされる。しかし、これだけでは、当該業務そのものの知識が獲

得可能だが、人間はこれに加えて対象やその操作に関する知識を前提として持っている。この部分はオントロジーとして別途補うようにしている。この場合のオントロジーは文書中の概念がどう互いに関係しているか程度の内容を持つオントロジーであり、具体的には重要な用語の切り出しと相互の関係の指示程度で構築可能であり、専門家であれば容易に可能なものである。すなわち、本研究の方法では、ほとんどの知識の整理は専門家が文書作成程度の労力で可能であり、この点でこれまでの方法とは異なっているといえる。

5.4 第5章のまとめ

本章では、オントロジー、インターネットロボット、エージェント, テキスト分類、内容処理および工学的知識の共有などの観点から関連研究と比較し、本研究の特色について述べた。また、オントロジーに基づく情報共有のアプローチの特色についていくつかの疑問点に答えていく形式で総括的な議論を行なった。

第6章

結論

本研の究目的は、オントロジーを用いた知的情報共有を実現するための方法論を明らかにすることである。

広範囲な情報の関連性を知るためには単に情報源の情報を利用するだけでは難しく、背景的知識が必要になる。オントロジーは、「対象領域に関する概念と概念間の関係の記述」と定義され、知的情報共有の鍵となる概念である。本研究では、その背景的知識を提供する語彙体系、概念体系としてオントロジーの利用の利用法、構築法について議論してきた。

具体的には

1. オントロジーを利用したネットワークからの情報の収集・分類・統合化法
2. オントロジーを利用した工学的知識の組織化・共有化法
3. オントロジーの構築支援方法

の3点について議論を行ない、本研究で手法の有効性を検証するために、プロトタイプシステムを構築し、ネットワーク上のデータを用いた評価実験と考察を行なった。

1.では、WWWに代表されるような多様性(形式面、内容面)、分散性、大規模性を扱う情報源群から、ユーザが必要な情報を収集し、整理するためのオントロジーの利用方法について議論した。具体的には、オントロジーによる情報収集・分類・抽出システムIICAを実装し、その有効性を検証するために、WWWにおける評価実験を行った。その結果、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・統合化を一貫して行なうことができ、ネットワーク上の広範で不均質な情報源を利用可能にすることがわかった。

2. については、より背景知識の必要な技術情報の共有化に焦点を当てた。すなわち、ベテランの技術者が個人的に持っている背景知識をいかに明示的に記述し、それを中心に技術情報をいかに構造化・共有するかがテーマであった。具体的には、ベテラン技術者が持っている経験やノウハウを後輩に継承するための枠組としてICoBの考え方を導入し、変圧器改修計画業務支援を事例とした現場技術共有システムOnTheSpotを実現した。また、その有効性を検証するために、現場技術者による評価を行なった。オントロジーを用いることで、対象に関して全般的な知識を付加することができ、ドキュメントを対象の性質や構造などによって構造化して提示することで、当該の業務遂行に必要な背景的な知識を学習したり、対象の変化や追加に即したドキュメントを修正したりする際に有効になるとがわかった。

3. については、利用面まで考慮に入れた、オントロジーの構築支援について焦点を

あてた。具体的には、オントロジーの獲得を (1) 概念の獲得、(2) 概念間の関係の獲得に分け、WWWのリンク情報、共起関係、情報量等を用いたオントロジー獲得支援法について、WWWページを用いた分類実験をもとに比較検討し、オントジーの自動獲得がどこまで可能性かについて考察した。その結果、提案する方法によって、オントロジーを手動で作成した場合と同等以上の分類精度が得られ、オントロジー作成の手間を短縮できることがわかった。

これらの結果から、オントロジーに基づくアプローチが、定義や構造が不明確で大規模な情報を、内容レベルで統合・共有化するのに有効な手段であることがわかった。

謝辞

主指導教官の西田豊明教授には、奈良先端科学技術大学院大学情報科学研究科博士前期過程および博士後期課程在学中の 5 年半の長きにわたりご指導を賜りました。先生には、研究の道すじを示していただき、研究内容や進め方を親身になってご指導いただいただけではなく、研究者としてのあり方をご教示いただきました。心より感謝致します。

武田英明助教授にもまた、副指導教官として 5 年半にわたりご指導いただきました。研究の要所で貴重な助言をいただきましたことを深く感謝致します。

松本裕治教授には、お忙しい中副指導教官となっていただき、また、審査委員もお引き受け下さいました。植村俊亮教授もまた、ご多忙の中、審査委員に加わって下さいました。お二人に的確なご指摘と有益なコメントをいただきましたことを感謝致します。

沢田篤史助手 (現在京都大学大型計算機センター助教授)、上野敦志助手、久米出助手には、研究会などでいろいろな意見をいただき、また開発環境の整備などで助けていただき、たいへん感謝しております。

西田研究室の方々には、公私にわたりたいへんお世話になりました。鷹合基行君、寺田和憲君には、研究内容だけでなく、大学生活全般にわたってアドバイスをいただきました。前田晴美さん (現在大阪市立大学学術情報総合センター講師)、糀谷和人君 (現在オムロン株式会社)、畑谷和右君 (現在松下電器産業株式会社)、白神謙吾君 (現在三菱コントロールソフトウェア株式会社)、大杉英一君 (現在 NTT マルチメディアネットワーク研究所) には、研究のアイディアに関する多くの有益な議論とシステムの開発、整備などでお世話になりました。事務補佐員の小布施文代さん、谷村優香里さんには、いつも迅速、的確に事務処理を行っていただき、研究生活を進める上でたいへん助かりました。

また、親友の竹内孔一君 (現在学術情報センター)、沖本忠久君 (現在 NTT-TE 関西営業本部)、康村昌司君 (現在日本総合研究所) には、公私にわたり大変お世話になり、時として単調になりがちな大学生活を充実したものにすることができました。ここに心から感謝いたします。

紙面の都合でここに名前をあげられなかった方々にも、学生生活や研究を進め

る上で有形無形にたいへんお世話になりました。深く感謝致します。

現場技術共有支援システムに関する研究では、関西電力株式会社総合技術研究所の水上雄一主席研究員、東光精機株式会社の高岡良行課長、太田衛氏、浅野耕作氏に、研究のアイディアに関する多くの有益なご意見とシステムの開発、整備などで大変お世話になりました。

また、関西電力 (株) の尾田純二氏には、改修設計に関する専門知識の提供およびプロトタイプシステムの評価をしていただいたに深く感謝したします．変圧器設備改修に関する資料・データの収集に協力していただいた関西電力 (株) 三田制御所の方々に心から感謝いたします．

最後になりましたが、研究生活を支えてくれた家族に深く感謝致します。

参考文献

[1] Thomas R Gruber. A translation approach to portable ontology specifications. In R. Mizoguchi, H. Motoda, J. Boose, Gaines B., and Quinlan R., editors, *Proceedings of the Second Japanese Knowledge Acquisition for Knowledge-based Systems Workshop JKAW '92*, pp. 89–108, 1992.

[2] Riichiro Mizoguchi. Task ontology and its use in a task analysis interview system. In *Proceedings of JKAW '92*, 1992.

[3] T. R. Gruber. Ontolingua: A mechanism to support portable ontologies. Technical Report KSL 91-66, Stanford University, Knowledge Systems Laboratory, 1992.

[4] 松尾文碩, 柴田誠, 竹田正幸. 科学技術用語オントロジーの自動生成. 情報学基礎,38-3, pp. 17–24, 1995.

[5] 長尾真, 石田晴, 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.

[6] 松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾真. 日本語形態素解析システム JUMAN 使用説明書 version 2.0. 1993.

[7] G.Salton and M.J.McGill. Introduction to Modern Information Retrieval. In *McGraw-Hill*, 1983.

[8] 林良彦, 菊井玄一郎, 鷲崎誠司, 砂場倫太郎. WWW 情報空間における Resource Discovery と Navigation 支援. 信学技法, No. AI95-31, 1995.

[9] 西田豊明, 武田英明. 知識コミュニティプロジェクト (第 4 報) —統合的知識環境を目指して—. 第 11 回人工知能学会全国大会, pp. 336–339, 1997.

[10] 溝口理一郎. オントロジー工学への道. 人工知能学会誌, Vol. 13, No. 1, pp. 9–10, 1998.

[11] Microsoft. Microsoft agent.
http://www.microsoft.co.jp/intdev/agent/.

[12] R. V. Guha and D. B. Lenat. Cyc: A midterm report. *AI magazine*, pp. 32–59, Fall 1990.

[13] Sharable ontology library.
http://www-ksl.stanford.edu/knowledge-sharing/ontologies/index.html.

[14] Yahoo!
http://www.yahoo.com/.

[15] Michael Pazzani,Jack Muramatsu,Daniel Billsus. Syskill & Webert: Identifying web sites. In *AAAI-96/IAAI-96 Proceedings Volume One*, pp. 54–61, 1996.

[16] 高岡良行, 広部健治, 溝口理一郎. 再利用可能知識ベースの構築 —変電所事故復旧問題を例にし て—. 人工知能学会誌, Vol. 10, No. 5, pp. 786–797, 1995.

[17] O. McBryan94. Genvl and WWWW: Tools for taming the web. In *Proc. 1st Int. WWW Conf.*, 1994.

[18] P. Maes. Agents that reduce work and information overload. *CACM*, Vol. 37, No. 7, pp. 30–40, 1994.

[19] M. Balabanovic and Y. Shoham. Learning information retrieval agents: Experiments with automated web browsing. In *Proc. AAAI Spring Symposium*, pp. 13–18, 1995.

[20] 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. 情処研報 R94012, 第 95 巻, pp. 7–12, 1995.

[21] 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.

[22] 亀田弘之, 藤崎博也. テーマ・キー概念・キーワード間の階層構造を利用する新聞記事情報の分類・検索システム. 情報処理学会論文誌, Vol. 28, No. 11, 1987.

[23] T. Kohonen. The self-organizing map. In *Proc. IEEE*, Vol. 78, pp. 1464–1480, 1990.

[24] 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.

[25] 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.

[26] 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.

[27] 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, 第 95 巻, pp. 87–94, 1995.

[28] D.K. Harman. The darpa tipster project. *ACM SIGIR FORUM*, 1992.

[29] D.E.Appelt, J.R.Hobbs, J.Bear, D.Israeland, M.Tyson. A finite-state processor for information extraction from real-world text. Proc. of 13th International Joint Conference on Artificial Intelligence, Vol.2, Morgan Kaufmann, pp. 137–140, 1993.

[30] 田中秀雄, 植田孝夫, 西田省吾. 共同作業による知識の創造, 蓄積, 伝承を支援する対話型教育支援システム. 電気学会論文雑誌 B, Vol. 111, No. 12, pp. 1455–1461, 1995.

[31] 高橋丈博, 渋谷昇, 伊藤健一. 事例を用いたノイズ理解と対策支援ツールの開発. 電気学会論文雑誌 C, Vol. 113, No. 8, pp. 591–597, 1993.

[32] 大嶽能久, 笹氣光一, 福井美佳, 後藤和之, 中山康子, 竹林洋一. 知識ベースとノウハウベースの連携による知識情報共有システムの実現. 第 11 回人工知能学会全国大会論文集, pp. 310–311, 1997.

[33] 溝口理一郎, 池田満. オントロジー工学序説 —内容指向研究の基盤技術と理論の確立を目指して—. 人工知能学会誌, Vol. 12, No. 4, pp. 559–569, 1997.

研究発表一覧

1. 著書 (分担執筆)

[1] Toyoaki Nishida, Hideaki Takeda, Michiaki Iwazume, Harumi Maeda, and Motoyuki Takaai. Chapter 5. The Knowledgeable Community: Facilitating Human Knowledge Sharing, in Communityware: Towards Global Collaboration, John Wiley & Sons , 1998.

[2] 西田豊明, 武田英明, 岩爪道昭, 前田春美, 鷹合基行. 知識コミュニティ: 第3章メディアに重点をおいたアプローチ, 朝倉 AI ライブラリ 6 工学知識のマネージメント, 朝倉書店, 1998(出版予定)

2. 論文誌

[1] 岩爪道昭, 白神謙吾, 畑谷和右, 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・分類・統合化. 情報処理学会論文誌, 38(3):606-615, 1997.

[2] 岩爪道昭, 武田英明, 西田豊明, 浅野耕作, 太田衛, 高岡良行, 水上雄一. オントロジーを用いた変圧器改修計画業務支援ドキュメントベースシステム. 電気学会論文誌, 1997. (条件付き採録)

3. 国際会議 (査読付き)

[1] Michiaki Iwazume, Hideaki Takeda, and Toyoaki Nishida. Ontology-based approach to information gathering and text categorization. In Proceedings of International Symposium on Didital Libraries 1995, pp.186-193, 1995.

[2] Michiaki Iwazume, Hideaki Takeda, and Toyoaki Nishida. Ontology-based information gathering and text categorization from the internet. In Takushi Tanaka, Setsuo Ohsuga, and Moonis Ali, editors, Proceedings of the Ninth International Conference in Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE-96), pp.305-314, 1996.

[3] Michiaki Iwazume, Hideaki Takeda, and Toyoaki Nishida. Ontology-based information capturing from the internet. In Proceedings of the fourth International Conference on the International Society of Knowledge Organization, pp.261-272, 1996.

[4] Michiaki Iwazume, Kengo Shirakami, Kazuaki Hatadani, Hideaki Takeda, and Toyoaki Nishida. IICA: An ontology-based internet navigation system. In Working notes for AAAI96 Workshop on Internet-Based Information Systems, pp.65-71, 1996.

[5] Yoshiyuki Takaoka, Kousaku Asano, Mamoru Ohta, Michiaki Iwazume, Hideaki Takeda, Toyoaki Nishida, and Yuuichi Mizukami. A Planning Support Document Base System for Transformer Reparation Task Using Ontology. In Proceedings of Second International Conference on Knowledge-Based Intelligent Electronic System 1998, pp.35-44, 1997.

[6] Toyoaki Nishida, Hideaki Takeda, Michiaki Iwazume, Harumi Maeda and Motoyuki Takaai. The Knowledgeable Community – Facilitating the Knowledge Process by Humans and Computers –. In Proceedings of Second International Conference on Knowledge-Based Intelligent Electronic System 1998, pp.23-32, 1997.

4. 国内会議 (査読付き)

[1] 岩爪道昭, 白神謙吾, 畑谷和右, 武田英明, 西田豊明. 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. 第 7 回データ工学ワークショップ, 1996.

5. 研究会等 (査読なし)

[1] 岩爪道昭, 武田英明, 西田豊明. 電子掲示板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.

[2] 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収集・分類のアプローチ. 人工知能学会全国大会 (第 9 回) 論文集, pp.387-390, 1995.

[3] 岩爪道昭, 武田英明, 西田豊明. 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. 電子情報通信学会技術研究報告, AI95-32, pp.63-70, 1995.

[4] 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, 第 95 巻, pp.25-32, 1995.

[5] 岩爪道昭, 武田英明, 西田豊明. オントロジーを用いた広域ネットワークからの情報獲得. 第 11 回ヒューマンインタフェースシンポジウム論文集, pp.95-100 , 1995.

[6] 岩爪道昭, 白神謙吾, 畑谷和右, 武田英明, 西田豊明. 知識に基づくインターネットからの情報獲得と統合化. 情報処理学会研究報告 96-FI-42, pp.17-22 , 1996.

[7] 太田衛, 高岡良行, 岩爪道昭, 武田英明, 西田豊明, 水上雄一. イントラネットによる現場技術情報共有化システム. 平成 9 年電気学会全国大会, 3, pp.107, 1997.

[8] 黒田直志, 大杉英一, 岩爪道昭, 武田英明, 西田豊明. ネットワークの情報を利用した概念獲得支援. 人工知能学会全国大会 (第 11 回) 論文集, pp.555-558 , 1997.

[9] 岩爪道昭, 武田英明, 西田豊明, 太田衛, 高岡良行, 水上雄一. 現場技術情報共有のためのオントロジーの作成と利用. 人工知能学会全国大会 (第 11 回) 論文集, pp.320-323, 1997.

[10] 岩爪道昭, 武田英明, 西田豊明, 太田衛, 高岡良行, 水上雄一. オントロジーを用いた現場技術情報共有の知的支援. 電子情報通信学会技術研究報告 OFS97-18, pp.33-40, 1997.

[11] 岩爪道昭, 武田英明, 西田豊明, 太田衛, 高岡良行, 水上雄一. 現場技術情報の体系化・共有のための知的ドキュメントベース,. 電子情報通信学会 第3回知能情報メディアシンポジウム 論文集, pp.317-314, 1997.

[12] 岩爪道昭, 武田英明, 西田豊明, 太田衛, 高岡良行, 水上雄一. 現場技術情報共有の知的支援. 第54回情報処理学会全国大会論文集, 第3巻, pp.295-296, 1997.

6. テクニカルレポート (分担執筆)

[1] Hideaki Takeda, Michiaki Iwazume, and Toyoaki Nishida. Ontology-centric knowledge organization. Technical Report NAIST-IS-TR97005, Nara Institute of Science and Technology, Nara, Japan, January 1997.

7. 授賞

[1] 第9回人工知能学会全国大会優秀論文賞, 1995.

付録

A. ヒューリスティックに基づく情報抽出ルール

2.6.3 項の実験で用いた、概念記述ルールを以下に示す。

A.1 2.6.3 項の実験で用いた記述ルール (寺)

```
(define-concept (寺の名前 (is "+寺" with (or h1 dt h2 h3 ))))

(define-concept (重要文化財 (is 歴史的遺物 with (or "重要文化財" "重文"))))

(define-concept (県指定文化財 (is 歴史的遺物 with (or "県指定文化財"))))

(define-concept (国宝 (is 歴史的遺物 with (国宝である))))

(define-concept (本尊 (is 仏像名 with ("本尊"))))
(define-concept (創建年 (is 年 with (創建))))

(define-concept (国宝である (or "国宝")))
(define-concept (寺宝 (is 歴史的遺物 with ("寺" "宝"))))

(define-concept (歴史的遺物 (or 仏像名 建物名 絵名)))
(define-concept (創建 (or "建立" "建てる" "建造" "造営" "創建" "建立"
                          "竣工" "創立")))
(define-concept (年 (or 日本暦（西暦） 日本暦 西暦 "年間")))
(define-concept (仏像名 (and (or "仏>" "天>" "尊>" "像>" "如来>" "菩薩>" "観音>" "明王>")
                             (not (or "天皇" "本尊")))))
(define-concept (建物名 (and (or "本堂" "堂" "門" "鐘楼" "金堂" "講堂" "塔"))))
(define-concept (絵名 (and (or "絵巻" "曼陀羅" ))))
(define-concept (日本暦 (and "年" (or 数字 "元"))))
(define-concept (日本暦（西暦） (list 日本暦 "（" 数字 "）" )))
(define-concept (西暦 (and "年" 数字)))

(define-concept (宗派 (or "宗>" (and "宗" "派>"))))

(define-concept (時代 (or "時代>" "年間>" 年)))


(define-concept (数字 (or "0" "1" "2" "3" "4" "5" "6" "7" "8" "9"
                          "０" "１" "２" "３" "４"
                          "５" "６" "７" "８" "９")))

(define-concept (大人料金 (and 料金 "大人")))
(define-concept (子供料金 (and (or "子供" "小人") 料金)))
(define-concept (大学生料金 (and "大学" 料金)))
(define-concept (高校生料金 (and "高校" 料金)))
(define-concept (中学生料金 (and "中学" 料金)))
(define-concept (小学生料金 (and "小学" 料金)))
(define-concept (学生料金 (and "学生" 料金 (not "大学生" "中学生" "小学生"))))
(define-concept (料金 (and 数字 "円")))
```

A.2 2.6.3 項の実験で用いた記述ルール (温泉)

```
(define-concept (最寄り駅 (is 駅 before (or "から" "より" "下車"))))
(define-concept (最寄り駅 (is 駅 with ("最寄り駅"))))
(define-concept (駅 (and "?駅名?" "駅")))
(define-concept (駅 (is "*名詞*" after (or "地下鉄" "名鉄" "+線>"))))
(define-concept (駅 (is "*名詞*" before (and "$格助詞$" (or "から" "より")))))
(define-concept (駅 (is "*名詞*" before ("下車"))))

(define-concept (アクセス方法 (is "徒歩" "分")))
(define-concept (アクセス方法 (is "バス" "分")))
(define-concept (アクセス方法 (is "下車")))

(define-concept (温泉の名前 (is 温泉名 2 by (or h1 dt h2 h3))))
(define-concept (温泉名 2 (or (list "<) >" "温泉>") "+温泉>")))

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

(define-concept (風呂の種類 (is "+風呂>")))
(define-concept (風呂の種類 (list "*" "の" "<湯>")))
(define-concept (風呂の種類 (and "+湯>" (not "開湯")
                    (not "入湯") (not "銭湯"))))


(define-concept (泉質 (is 泉の名前 with (or "泉質>" "含む"))))
(define-concept (泉の名前 (or (and "+泉>" (not "<温泉>"))
                    "単純")))

(define-concept (電話番号 (list 数字 バー 数字)))
(define-concept (バー (or "-" "ー" "ー")))
```

A.3　2.6.3 項の実験で用いた記述ルール (飲食店 1)

```
(define-concept (飲食店の名前 (is (or "+屋>" "+軒>") with (or h1 dt h2 h3))))

(define-concept (料理の種類 (or (list 食物 "の" (or "料理>" "店>"))
                    (and 食物 (or "+屋さん>" "+屋>"))
                    "+料理>" "<中華>" "<洋食>" "<和食>"  "居酒屋>"
                    (and 食物 "+類>") "+飯店>" "飲み屋>"
                    "+レストラン>" "<定食>" "喫茶>" "喫茶店"
                    )))

(define-concept (おいしさ (is 食物 with (or "おいしい" "まずい" "まあまあ"
                        "美味しい" "ほちほち" "ボチボチ"
                        "うまい"
                        ))))


(define-concept (甘さ (is 食物 with (or "甘い" "辛い" "しょっぱい" "苦い"
                        "にがい"))))

(define-concept (食物 (or "うどん" "定食" "中華そば" "ラーメン" "コーヒー"
                    "カレーシチュー" "オムレツ" "クッキー" "カキ"
                    (list "<たこ>" "<焼く>") "カレー" "チャーシュー麺"
                    "チャーシュウ麺" "チャンポン" "チャーシューメン"
                    "ギョウザ" "弁当" "チャーハン" "餃子"
                    (list "<お>" "<好む>" (or "<焼く>" "<焼>"))
                    "焼肉" "そば焼" "ぶた焼" "ミックス焼" "塩タン"
                    "クッパ" "ビビンバ" "冷麺" "ケーキ" "ガム" "キムチ"
                    "チーズ焼" "ずり焼" "<ピザ焼>" "そば" "サラミ"
                    "うなぎ" "ぎょうざ" "中華丼" "天津丼" "牛丼"
                    "おにぎり" "いなりずし" "肉じゃが" "空揚げ"
                    "釜飯" "ピンクしゃわしゃわ" "魚" "トンカツ"
                    "カツ丼" "かつ丼" "ご飯" "ごはん" "パフェ" "ピザ"
                    "枝豆" "焼き鳥" "にくじゃが" "もつ煮" "はまち刺"
                    "焼き串" "コロッケ" "バーベキュ" "ハヤシライス"
                    "イディワッパ" "スパゲティ" "スパゲッティ"
                    "豚生セット" "たぬきセット" "<ライス>" "すきやき"
                    "すき焼き" "ステーキ" "シーフード" "ストーンクラブ"
                    "ハンバーグ" "リゾット" "パン>" "パスタ" "炒飯"
                    "サラダ" "チキンカツ" "ドリアセット" "ドリア>"
                    "パンドミ" "アイスクリーム" "+餅" "+麺>"
                    "味噌汁" "たくあん" "冷奴" "+弁当>" "串かつ"
                    "とんかつ" "<タラ>" "<デザート>" "マーボーライス"
                    "+ランチ>" (list "*" "の" "<ランチ>") "エビちり"
                    "生ちらし" "すし>" "寿司>" "にぎり" "刺身"
                    "+いため>" "+炒め>" "おしんこ" "漬物" "ねぎとろ丼"
                    "ヨーグルト" "ポークソテー" "シャーベット"
                    "コーンフレーク" "ロールキャベツ" "ソーセージ"
                    "メンチカツ" "炒め野菜" "肉野菜丼" "有明ボンメン"
                    "ラーシーメン" "焼き肉" "ウニ" "山菜" "あんみつ>"
                    )))

(define-concept (汁 (or "スープ>"
                    )))

(define-concept (飲物 (or "ビール>" "酒>" "ワイン>" "ジュース>" "カクテル>"
                    "コーラ>" "ヤムチャ" "飲茶" "紅茶" "コーヒー>"
                    "+茶>" "+ティ>"
                    )))
```

A.4　2.6.3 項の実験で用いた記述ルール (飲食店 2)

```
(define-concept (素材 (or "たまねぎ>" "にんにく>" "チーズ" "ねぎ>" "葱"
                "えび天" "エビ" "山菜" "天かす" "てんかす" "+ソース>"
                "キャベツ" "チャーシュー" "<めんま>" "メンマ"
                "<もやし>" "コーン" "バター" "しなちく" "昆布"
                "ごま>" "ニンジン" "にんじん>" "人参" "<しめじ>"
                "海苔" "<のり>" "ノリ""ひきにく" "挽き肉" "<わかめ>"
                "引き肉" "わかめ>" "若布>" "ハム>" "<ふ>" "とり肉"
                "たまご>" "卵>" "マロン" "栗" "<かつお" "<まぐろ>"
                "<いか>" "イカ" "こはだ" "ばい貝" "<たこ>" "<あじ>"
                "マヨネーズ" "グリーンピース" "茄子" "<なす>"
                "キクラゲ" "<アサリ>" "ホウレンソウ" "<ソース>"
                "かまぼこ>" "ちくわ>" "半チャンセット" "+タマゴ"
                )))

(define-concept (材料 (is 素材 with (or "入る" "かける" "山盛り" "のっかる"
                        (list "ベース" "に") "かかる"
                        ))))

(define-concept (味 (is 味覚 with (and "味" (not "味わえる") (not 美味しさ)))))
(define-concept (味 (and 味覚 "系")))

(define-concept (味覚 (and (or "濃い" "薄い" "塩" "醤油" "しょうゆ" "みそ"
                  "味噌" "不思議だ" "とんこつ" "豚骨" "あっさり"
                  "サラサラ" "こってり" "コッテリ" "砂糖"
                  (list "うらぎり" "の" (or "ある" "ない"))
                  "落ちる" "上品だ" "しつこい" "アッサリ"
                  "珍しい"
                  )
                 (not 食物))))

(define-concept (値段 (is 安さ with (or "お金" "値段" "料金" 食物))))
(define-concept (値段 (and 数字 (or "+円>" "<¥+"))))
(define-concept (値段 (is "<タ ダ>")))

(define-concept (安さ (or "高い" "安い" "まあまあ" "高め")))

(define-concept (やわらかさ (is 食物 with (or "ふっくら" "硬い" "かたい"
                          "やわらかい" "柔らかい"
                          "ふわふわ" "カリカリ"
                          (list "ギトギト" "する")
                          ))))
(define-concept (やわらかさ (or (list "粘り" "の" (or "ない" "ある"))
                  with ("麺"))))

(define-concept (年齢層 (is "+向き>")))

(define-concept (お勧め (is 食物 with (or "有名" "おすすめ" "楽しめる"
                        (list "<価値>" "<は>" "ある")
                        ))))

(define-concept (調理方法 (is (or 焼き 煮る) with (食物))))
(define-concept (焼き (is (or "二度") with ("焼く"))))
(define-concept (煮る (is "煮る")))
```

A.5　2.6.3 項の実験で用いた概念のスロット構造

```
(define-pclass (飲食店 (
                (has-one 飲食店の名前)
                (has-some 料理の種類)
                (has-some やわらかさ)
                (has-some おいしさ)
                (has-some 甘さ)
                (has-some 味)
                (has-some 値段)
                (has-some 材料)
                (has-some 年齢層)
                (has-some お勧め)
                (has-some 調理方法)
                )))

(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)
                )))

(define-pclass (訪問地 (
                 (has-one 最寄り駅)
                 (has-one 駅からのアクセス方法)
                 (has-one アクセス方法)
                 (has-one 電話番号)
                 )))

(define-pclass (有料訪問地
          (
           (is-a 訪問地)
           (has-some 料金)
           (has-one 大人料金)
           (has-one 学生料金)
           (has-one 子供料金)
           (has-one 大学生料金)
           (has-one 高校生料金)
           (has-one 中学生料金)
           (has-one 小学生料金)
           )))

(define-pclass (寺 ((has-one 寺の名前)
                (is-a 有料訪問地)
                (has-one 創建年)
                (has-some 国宝)
                (has-some 重要文化財)
                (has-some 県指定文化財)
                (has-one 本尊)
                (has-some 寺宝)
                (has-one 宗派)
                (has-some 時代)
                )))
```