連想構造を用いた情報整理システム

前田晴美[†] 糀谷和人[†,☆] 西田豊明[†]

既存の雑多で構造の不均質な情報源から情報を収集・整理する手法を提案する．基本となるアイデアとして，雑多な情報をゆるやかに関連づける連想構造というデータ構造を用いる．連想構造は生データから容易に生成でき，人間が直観的に理解しやすい．我々は，連想構造を用いて既存の情報源から情報を収集し，整理する過程を支援するシステム CM-2（Contextual Media version 2）を試作した．CM-2 では，(a) 既存の情報源から情報を取り込み，連想構造を生成する情報キャプチャ機構と，(b) キーワードに基づき情報の切り出しと構造化を行う知的情報統合機構を実現した．CM-2 の有効性を実験によって確かめた．

Information Reorganization Using Associative Structures

HARUMI MAEDA,[†] KAZUTO KOUJITANI[†,☆] and TOYOAKI NISHIDA[†]

We propose a method to gather and reorganize information from heterogeneous information sources. The method is based on the use of a plain information representation called *associative structures*. *Associative structures* connect various information media without defining the semantics rigorously. They are easy to generate from raw data and comprehensible to humans intuitively. We developed a system called CM-2 (Contextual Media version 2) to realize this method. We describe the system's two major facilities; (a) an *information capture facility* which gathers information from heterogeneous information sources and generates associative structures and (b) an *intelligent information integration facility* which reorganizes information according to user's input. We verify our approach by analyzing results of experiments.

1. はじめに

研究活動に代表される創造的思考活動には雑多で構造の不均質な情報源の情報が必要である．たとえば，過去に作成した論文や OHP や研究メモ，研究動向を調査するための書籍・雑誌，広く知識を収集するための辞書，オンラインデータベースや WWW（World Wide Web）などである．文書として書き表されていない頭の中の断片的なアイデアや常識も重要である．このような情報を統合的に整理するためには，既存のデータモデルや知識表現を用いることは人間の負荷が高い．

本論文では，既存の雑多で構造の不均質な情報源から情報を収集・整理する手法を提案する．基本となるアイデアとして，雑多な情報をゆるやかに関連づける**連想構造**というデータ構造を用いる．連想構造は生データから容易に生成でき，人間が直観的に理解しやすい．採用する方法はヒューリスティックなものであり，あらゆるケースに対応できるものではないが，予備実験の結果かなりの有効性が期待されたので，定式化，実現して有効性を実験的に評価することとした．我々は，連想構造を用いて既存の情報源から情報を収集し，整理する過程を支援するシステム CM-2☆☆を試作した．CM-2 では，

- 既存の情報源から情報を取り込み，連想構造を生成する**情報キャプチャ機構**
- キーワードに基づき情報の切り出しと構造化を行う**知的情報統合機構**

を実現した．

本論文は以下のとおり構成される．2 章では既存の情報源の雑多性について述べる．3 章では連想構造と

† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology
☆ 現在，オムロン株式会社新事業開発センタ
Presently with New Business Development Division Headquarters, OMRON Corporation

☆☆ "CM" は我々の長期的な研究目標である "Contextual Media（文脈メディア）" の略語である．

(a) James Allen のページ (b) Barbara Hayes-Roth のページ

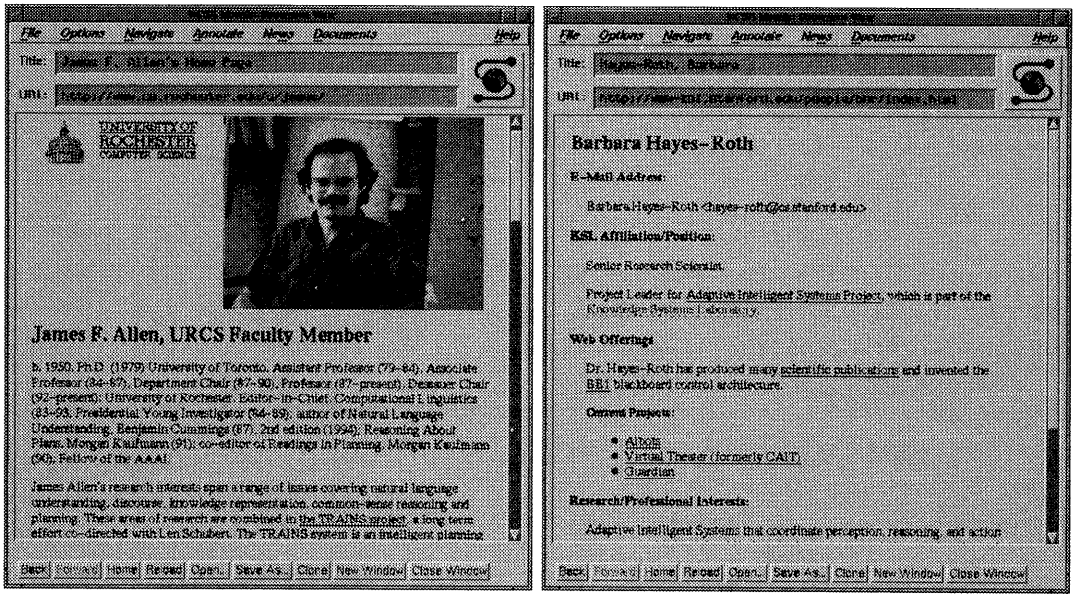

図1 WWW ページの例
Fig. 1 Example WWW pages.

CM-2 の概要について述べる．4 章では例題を示す．5, 6 章では CM-2 の 2 つの手法について説明する．7 章では実験結果を示す．8 章では手法の有効性について議論する．9 章では関連研究と比較する．

2. 既存の情報源の雑多性

我々の身の回りには多様な情報源が存在する．**表 1** に，情報源の例を示す．情報源は電子化されているか，いないかによって分類できる．電子化されているものは，テキスト，画像（静止画・動画）などに分類できる．テキストは，プレーンテキスト，見出し付きテキスト，構造化テキストなどに分類できる．このような形態の違いを**形態の雑多性**と呼ぶ．

概念にどのような名前をつけるかは人によって違う．違う名前で同じ概念の場合（同義）や，同じ名前で違う概念の場合（多義）がある．同じ人を指す場合に，姓のみ，すべての名前，ミドルネーム省略など，多様な書き方がある．また，日本語にはかな表記のゆれと呼ばれる現象がある[1]．これらの現象を**表記の雑多性**と呼ぶ．

同じ形態のテキストでも，1 つ 1 つのテキストの構造は異なっている．たとえば，**図 1** (a)☆ と図 1 (b)☆☆を比較してみる．これらは，WWW ブラウザに表示された人工知能研究者 2 人のページである．直観的に明らかなように両者のページの構造は異なっている．たとえば，James Allen のページ（図 1 (a)）では，ページの最初に写真や名前があり，次に履歴や研究に関する情報が自然な文章ととして書かれている．これに対して Barbara Hayes-Roth のページ（図 1 (b)）では，名前，E-Mail アドレス，肩書などが箇条書きで書かれている．これらの現象を**構造の雑多性**と呼ぶ．

表 1 情報源の例
Table 1 Example of information sources.

情報源	電子化	情報の形態
WWW	○	ハイパーテキスト，画像など
NetNews	○	見出し付きテキスト
E-Mail	○	見出し付きテキスト
新聞記事	△	見出し付きテキスト
文献	△	構造化テキスト
個人のメモ	△	プレーンテキスト
プログラム	○	構造化テキスト
OHP	△	画像
アイデア	×	非定型
本	×	構造化テキスト

○：電子化されている △：一部電子化されている
×：電子化されていない

3. 情報整理システム CM-2

3.1 連想構造

連想構造は key と呼ぶいくつかのユニットと value と呼ぶいくつかのユニットの間に定義され，「key が与

☆ http://www.cs.rochester.edu/u/james/
☆☆ http://www-ksl.stanford.edu/people/bhr/index.html

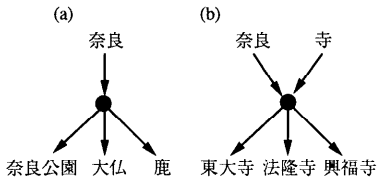

図2 連想構造
Fig. 2 Associative structures.

えられるとvalueが想起される」というゆるやかな関連を表す．keyとvalueの間の連想関係は厳密に定義されるのではなく，多分に主観的であることを許している．これは既存の情報源の雑多性に対応することをねらっている．

ユニットとは，情報を蓄積する情報ベースの基本構成要素である．ユニットはCM-2の外部のテキストファイルやイメージファイルを表す外部参照データユニットと，CM-2において情報を関連づけるための内部的な対象である概念ユニットに大別される．

本論文ではCM-2における連想構造を図2のように記述する．1つの連想構造を点を用いて表現し，keyとvalueの関係を矢印を用いて表す．図2(a)はkey「奈良」が与えられるとvalue「奈良公園」「大仏」「鹿」が想起されることを示す．図2(b)はkey「奈良」と「寺」からvalue「東大寺」「法隆寺」「興福寺」が想起されることを示す．

連想構造には，特別な形態として，IS-A構造，辞書構造を定義できる．IS-A構造では，クラス，サブクラスなどを区別することなく，上位概念をクラス，下位概念をインスタンスと呼ぶ．IS-A構造は知的情報統合機構における推論に使用される．辞書構造は概念ユニット間の変換に使用される．

3.2 CM-2の概要

CM-2は，既存の雑多で構造の不均質な情報源から情報を収集・整理するための情報整理システムである．

CM-2の情報ベースは，内部知識ベース（概念ベース）とテキストデータやイメージデータなどがある外部データベースから構成される．利用者はワークスペースと呼ぶ空間を通してネットワークに接続された個人やグループの情報ベースを利用できる．

CM-2では，(a)既存の情報源から情報を取り込み，CM-2の連想構造を生成する**情報キャプチャ機構**，(b)キーワードに基づき情報の切り出しと構造化を行う**知的情報統合機構**，の2つの機構を持つ．図3に，CM-2の処理の概要を示す．情報キャプチャ機構によって既存の情報源から情報を取り込んで連想構造を生成し，知的情報統合機構を用いて連想構造から情報を切り出して整理する．

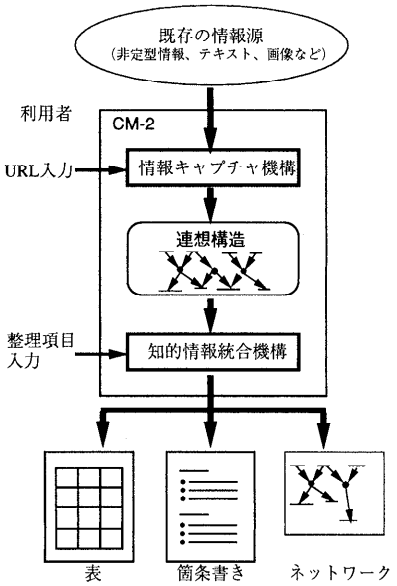

図3 CM-2の概要
Fig. 3 Overview of CM-2.

4. 例題

たとえば，興味のある分野の研究者やプロジェクトに関連する情報を集めたいが，適当な本やデータベースがないとする．そのような場合，個人の持つ知識や，WWWのページを統合して，必要な情報だけを整理できると便利である．

我々は，CM-2を用いて，図1のようなWWWの人工知能研究者のホームページを収集して，人工知能に関する情報を整理した．

4.1 例 1

図4(a)は，利用者が抽出キーワード「reasoning」，整理項目「研究者」「e-mail」「プロジェクト」「大学」を入力して表形式を選択して，研究者ごとに情報を整理した結果を示す．たとえば，「reasoning」というキーワードがページの中に書かれている研究者として，Adam Farquhar，Alon Levy，James Allenなどが抽出されている．James Allenに関しては，James Allenのページから名前，大学，プロジェクトの情報を抽出して表示している．元のページに書かれていないe-mailなどの情報は表示されない．

4.2 例 2

図4(b)は，整理項目「プロジェクト」「研究者」「e-mail」「大学」を入力，箇条書き形式を選択して，プロジェクトごとに情報を整理した結果を示す．たとえば，Adaptive Intelligent Systemsプロジェクトは，Bar-

(a) 研究者ごと（表）

抽出キーワード:「reasoning」

研究者	e-mail	プロジェクト	大学
Adam Farquhar	■ Adam_Farquhar@ksl.stanford.edu	■ PTTP	■ University of Texas at Austin ■ Stanford University
Alon Y. Levy	■ levy@research.att.com		■ Hebrew University ■ Stanford University
Brian Falkenhainer			■ MIT ■ University of Illinois at Urbana-Champaign
David Brown	■ sauron@wpi.edu ■ webmaster@cs.wpi.edu		■ Ohio State University ■ University of Kent ■ Michigan State University ■ USC
Edward A. Feigenbaum		■ Heuristic Programming Project	■ Carnegie-Mellon ■ National University of Singapore ■ Aston University
James F. Allen		■ TRAINS	■ University of Rochester ■ University of Toronto
Janet Kolodner		■ EXPEDITOR ■ MEDIC ■ CELIA	■ Yale ■ Brandeis University

(b) プロジェクトごと（箇条書き）

プロジェクト

Adaptive Intelligent Systems

- 研究者: Barbara Hayes-Roth, David Ash, Lee Brownston, John A. Drakopoulos, Philippe Morignot, Rich Washington
- e-mail: morignot@ksl.stanford.edu

an Undergraduate Computer Networks

- 研究者: David Finkel

BASAR

- 研究者: Christoph G. Thomas

Berkeley UNIX Consultant

- 研究者: Robert Wilensky, J. Martin

CABINS

- 研究者: Katia Sycara

CADET

- 研究者: Katia Sycara

CADIS

- 研究者: Katia Sycara

CAIT

- 研究者: Barbara Hayes-Roth

CAMIS

- 研究者: Torsten Heycke

CASE

(c) 研究者ごと（ネットワーク）

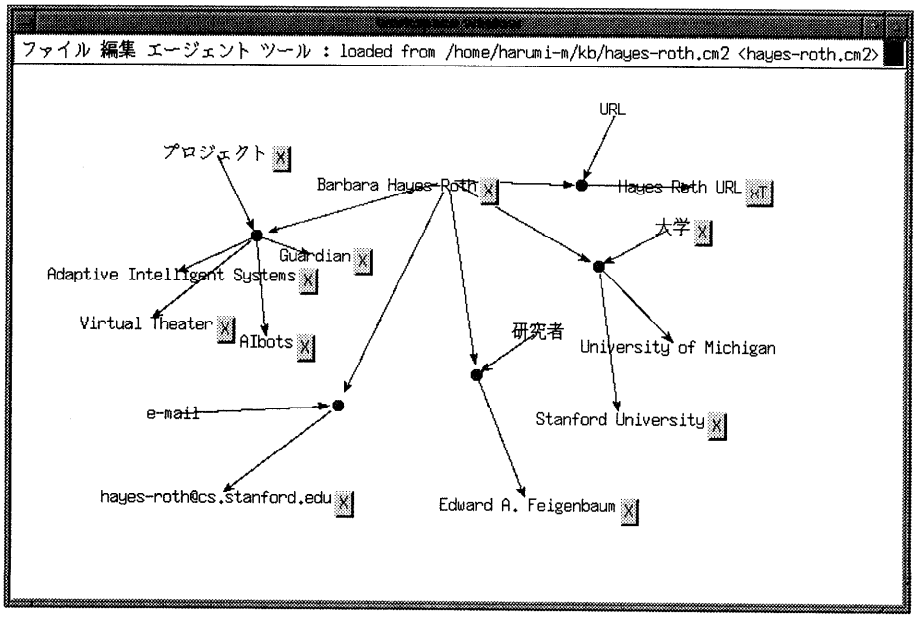

図4 CM-2の整理結果例
Fig. 4 Example results of CM-2.

bara Hayes-Roth のページや，Philippe Morignot のページ☆から抽出され，プロジェクトから逆に研究者名を関連づけて表示している．

4.3 例 3

図 4 (c) は，整理項目「研究者」「研究者」「URL」「e-mail」「プロジェクト」「大学」を入力，ネットワーク形式を選択した場合の結果の一部で，Barbara Hayes-Roth に関するワークスペースを示す．ワークスペースの中に，ドット表記に基づくネットワークによりユニットと連想構造が表示される．たとえば，key「Barbara Hayes-Roth」「研究者」から value「Edward A. Feigenbaum」が示されている．これは Feigenbaum のページに，Barbara Hayes-Roth の名前が書かれていることによる．このように，最初の整理項目と同じ項目を入力することにより，同一クラスのインスタンス間の関係を調べることができる．

ワークスペース上でユニットを選択することにより，連想されるユニット（value）を表示することができる[2]．また，ワークスペース上で，ユニットや連想構造の編集ができる[3]．

以下では，CM-2 の 2 つの機構の実機能について述べる．

5. 情報キャプチャ機構

情報キャプチャ機構は，既存の情報源から情報を取り込み，連想構造を生成する．

我々は，簡単なテキスト解析アルゴリズムとヒューリスティックを用いて，WWW 上の HTML 文書，日本経済新聞全文記事データベース（以下日経 DB と略す），Lisp プログラム，UNIX ファイルシステムなどを対象に，連想構造を生成するプログラムを開発した．

また，電子化されていない情報源から情報を取り込むために，ワークスペースから連想構造を直接入力できる．

5.1 アルゴリズム

アルゴリズムは情報源の種類によって異なるが，基本的には，**step 1** ユニットと連想構造の生成，**step 2** IS-A 構造の生成の 2 段階から構成される．以下，WWW 上の HTML 文書に焦点をあてて説明する．

我々は，利用者にドメイン知識がない場合や，最初にどんな情報がほしいか分からない場合も考慮して，形態素解析と HTML 構造解析を用いて汎用的なキーワードを抽出してから，ヒューリスティックを利用して重要なユニットを抽出する方針とした．URL を解析して HTML 文書を取得した後のアルゴリズムの概略を以下に述べる．

step 1 ユニットと連想構造を生成する．

step 1.1 形態素解析を用いてユニットを生成する．

- 形態素解析は，日本語では JUMAN[4]，英語では Brill's Rule Based Tagger[5] を用いる．
- 抽出する概念ユニットは，名詞または，名詞が連続して出現する品詞群である．英語の場合は，間に前置詞や接続詞や冠詞を含んでもよい．

step 1.2 HTML の構造を解析して連想構造を生成する．

- <h1> や <h2> などの見出し表現，<dl>，<ul>， などの箇条書き表現がある場合などにタイトル–スコープを設定し，ユニットがどのタイトル–スコープに含まれているかを基に連想構造を生成する．

step 2 ヒューリスティックを用いて IS-A 構造を生成する．

step 2.1 クラス判定ヒューリスティックを用いて IS-A 構造を生成する．

- クラス判定ヒューリスティックは，people 判定や project 判定などの個別クラス判定ヒューリスティックから構成される．ユニットに対して，個別クラス判定ヒューリスティックを順番に適用し，最初の判定結果をそのユニットのクラスとする．個別クラス判定ヒューリスティックは，与えられたユニットの名前に含まれる文字列からクラスを推論する．

step 2.2 **step 1** で生成されたユニットから IS-A 構造で連結されないものを削除し，連想構造を修正する（オプション）．

- この処理は選択が可能で，情報ベース内のデータ量の削減を目的としている．

5.2 例

James Allen のホームページの URL が与えられた場合どのような処理を行うのか説明する（**図 5**）．

URL 解析により HTML ファイル取得後，形態素解析を用いて，名詞句とそれに準じるものを概念ユニットとして抽出する．見出し表現<h1>...</h1> に囲まれる範囲から生成されたユニット「James F.Allen」

☆ http://www-ksl.stanford.edu/people/morignot/index.html

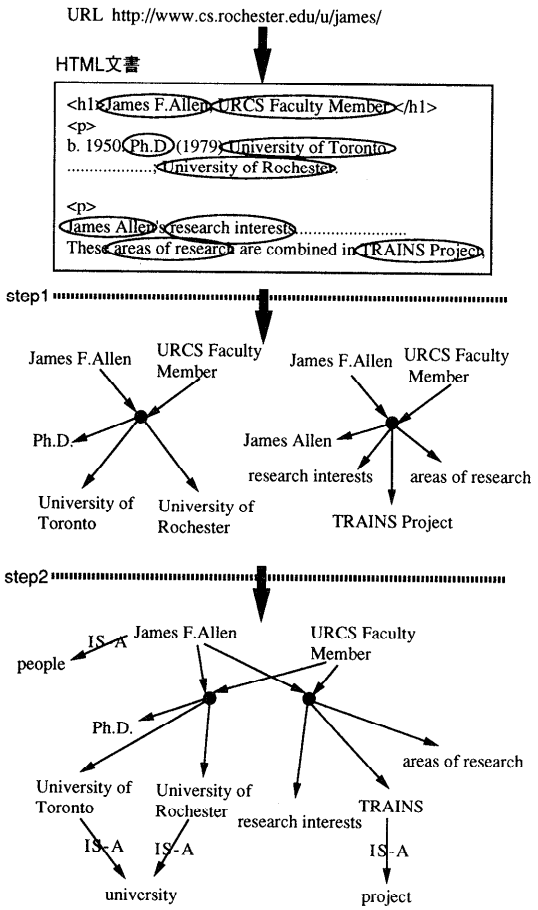

図5 情報キャプチャ機構のアルゴリズム動作例
Fig. 5 Information capture facility.

「URCS Faculty Member」を key とし，段落表現 <p>...<p> で囲まれる範囲から生成されたユニット「Ph.D.」「University of Toronto」「University of Rochester」を value とする連想構造を生成する（step 1）.

「James F.Allen」は「James」という英語名の単語を含むため「people」クラスであると推論し，IS-A 構造を生成する.「University of Rochester」「University of Toronto」は「university」という単語を含むため「university」クラスであると推論し，IS-A 構造を生成する（step 2）.

6. 知的情報統合機構

情報キャプチャ機構により生成された生の連想構造は，そのままでもワークスペース上で検索することができるが，構造や表記の雑多性が残る．知的情報統合機構は，構造や表記の雑多性を吸収し，不均質な連想構造から，利用者の入力したキーワードに基づき，情報の切り出しと構造化を行う.

6.1 アルゴリズム

アルゴリズムの概略を以下に示す．基本的には，**step 1** ヒューリスティックを用いた連想構造の統合，**step 2** 利用者の入力に基づく均質な連想構造の生成，**step 3** 結果の表示の3段階から構成される.

step 1 ヒューリスティックを用いて連想構造を統合する.

主要なヒューリスティックを以下に示す.

- **同名概念統合**：名前の文字列が同じ概念ユニットを統合する.
- **辞書参照概念統合**：辞書構造を参照して概念ユニットを統合する.
- **包含的連想生成**：ある概念ユニットの名前の文字列が他の概念ユニットの名前の文字列に含まれているとき，その概念ユニットを key として他の概念ユニットを value とする連想構造を生成する.
- **文脈参照連想統合**：ある連想構造と他の連想構造の key が同じとき，連想構造を統合する.
- **重複連想統合**：重複する連想を統合する.

このほか，表記の雑多性を解消するために，複数形統合ヒューリスティック（英語），ミドルネーム省略ヒューリスティック（英語），略語統合ヒューリスティック（英語）などを実装した.

step 2 利用者の入力に基づき均質な連想構造を生成する.

- 最初の整理項目をクラスとするインスタンスを最初の key として，2番目以降の整理項目を2番目の key として，複数のユニット間の連想構造の連鎖を探索する活性伝播型連想検索サブルーチン[6]を用いて value を求め，新しい連想構造を生成する.
- 抽出キーワードが入力されている場合は，最初の key を求める際に，インスタンスの集合の中から，抽出キーワードの文字列を含むユニットと連想構造で連結されているユニットを選択する.

step 3 生成された連想構造を表，箇条書き，ネットワーク形式に変換して表示する.

6.2 例

情報キャプチャ機構によって自動生成された Barbara Hayes-Roth と Phillipe Morignot に関する連想構造と，利用者が手入力した連想構造が与えられたときにどのような処理を行うのか説明する（**図6**）.

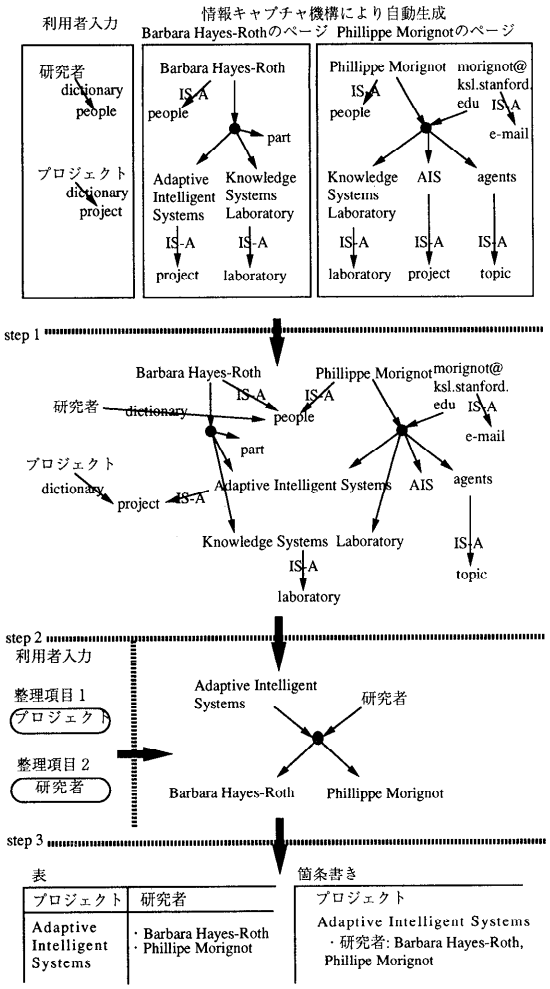

図6 知的情報統合機構のアルゴリズム動作例
Fig. 6 Intelligent information integration facility.

略語統合ヒューリスティックにより,「Adaptive Intelligent Systems」と「AIS」を統合する．同名概念統合ヒューリスティックにより,「Knowledge Systems Laboratory」「university」などを統合する（step 1）.

利用者が整理項目1「プロジェクト」，整理項目2「研究者」を入力すると,「プロジェクト」または「project」クラスのインスタンスであるユニット「Adaptive Intelligent Systems」を最初のkeyとする．項目2「研究者」を2番目のkeyとする．「研究者」または「people」クラスのインスタンスであるユニットの中から,「Adaptive Intelligent Systems」と連想構造でつながりがあるユニット「Barbara Hayes-Roth」と「Phillipe Morignot」をvalueとして，連想構造を生成する（step 2）.

7. 実　　験

CM-2を用いて異なる情報源から情報の収集・整理実験を行った.

7.1 実験1：奈良観光ガイドブック

奈良観光ガイドブック[7]を参考にして，奈良観光情報ベースを構築した．構築に際しては，ユニットと連想構造の記述を手作業で行ってから，IS-A構造の生成を行った．たとえばガイドブックに「東大寺の二月堂でお水取りが行われる」という情報がある場合，key「東大寺」「二月堂」value「お水取り」という連想構造を作成した．東大寺の写真（画像ファイル）がある場合には，key「東大寺」「写真」value「東大寺の写真（外部参照データユニット）」という連想構造を生成した．手作業で構築した情報ベースは，概念ユニット1276個，外部参照データユニット39個，連想構造861個の合計2176個である．IS-A構造に関しては，一部情報キャプチャ機構（step 2）を実装し，自動的に生成した．固有名詞辞書を与えずにどこまでできるかを実験した．たとえば,「寺」判定ヒューリスイックは,「寺」を文字列に含むかどうかの簡単な知識である.「寺」「神社」「値段」など26のクラスに関して実験したところ，350個のIS-A構造が生成され，そのうちの288個（82%）が適切であった．固有名詞辞書がなければIS-A構造の生成が難しいのは,「レストラン」「人」クラスなどである.

次に，自動生成されたIS-A構造に関連するクラスに関して，知的情報統合機構を用いて情報整理の実験を行った．適合率において,「寺」ごとに「花」「塔」「像」「駐車台数」の4つのクラスを整理するテスト1の結果が92%,「花」ごとに「寺」「時期」の2つのクラスを整理するテスト2の結果が86%であった（**表2**）☆.

連想構造を用いることにより，あらかじめ情報ベースのスキーマ設計を厳格に行わなくても，情報源の情報記述の流れに沿ってデータ入力ができることが分かった．また，ユニットが普通名詞である場合は，連想構造を生成したあとに，IS-A構造を自動生成することが容易でかつ有効であることが分かった.

7.2 実験2：日本経済新聞全文記事データベース

日経DBの記事は見出し付きテキストである．見出しはあらかじめ本文の内容を簡潔に表すように作成されており，その中に含まれる概念を抽出することにより，新聞記事の理解や加工を助けられると考える.

☆ 実験1では元の情報ベースを手作業で構築したので，再現率の評価は行っていない.

表2 実験結果
Table 2 Results of experiments.

情報源	テスト	適合率	再現率
奈良観光ガイドブック	テスト1	92%	−
	テスト2	86%	−
日本経済新聞全文記事データベース	テスト1	63%	91%
	テスト2	83%	71%
WWW上のHTML文書	テスト1	90%	83%
	テスト2	68%	73%

適合率：$\frac{\text{正しく生成された連想構造のユニット数}}{\text{生成された連想構造のユニット数}} \times 100$（%）

再現率：$\frac{\text{正しく生成された連想構造のユニット数}}{\text{生成されるべきユニット数}} \times 100$（%）

見出しから日本語形態素解析JUMANを用いて名詞を抽出して概念ユニット（key）として，記事本文を参照する外部参照データユニットをvalueとする連想構造を生成する情報キャプチャ機構（step 1）を実装した．1994年の日経DBの中で「インターネット」という単語を含む記事285件につき実験を行ったところ，概念ユニット1188個，連想構造285個，外部参照データユニット285個が生成された．生成された概念ユニットのうち92%が本文のキーワードとして適切であった．次に，実験1と同様に固有名詞辞書なしで情報キャプチャ機構（step 2）を実装し，「大学」「官公庁」クラスに関連するIS-A構造を生成した．

知的情報統合機構により，大学ごと（テスト1）および官公庁ごと（テスト2）に記事を整理する実験を行った．ここでは，単に組織名が含まれているというのではなく，その組織が記事の主体であるかという観点で評価を行った．適合率において大学ごとで適合率63%，再現率91%，官公庁ごとで適合率83%，再現率71%であった（表2）．

日経DBのような新聞記事データベースでは，見出しがついていることと，表記の雑多性が少ないことより，見出しに含まれる情報からだけでもある程度役に立つ整理ができることが分かった．

7.3 実験3：WWW上のHTML文書

4章の例題に関する実験を行った．人工知能研究者のホームページ100個から，情報キャプチャ機構を用いてHTML文書（英語）から情報を抽出し，知的情報統合機構を用いて7つの整理項目に関して整理を行った．

情報キャプチャ機構において，ユニット764個，連想構造625個，IS-A構造755個が生成された．これはデータ量の増加をおさえるためにstep 2において一部の連想構造とユニットを削除した数である．クラス判定ヒューリスティックのために与えた知識は，「people」「e-mail」「project」「university」「department」「laboratory」「topic」の7つのクラスに関して288個である．ただし，「e-mail」「university」「department」「laboratory」に関しては，実験1，2と同様に簡単な知識のみである．「project」に関しては，あらかじめ固有名詞辞書を与えずに，「文字列にprojectを含む」「大文字3文字以上である」などの知識を与えた．「people」「topic」に関しては，それぞれ英語の名前辞書，AI用語辞書の知識を与えた．

次に，知的情報統合機構を用いて，「people」ごとに他の6つのクラスを整理するテスト1と，「project」ごとに他の6つのクラスを整理するテスト2を行った．peopleごとで適合率90%，再現率83%，projectごとで適合率68%，再現率73%の確度で情報整理が正しく行われることを確認した（表2）．projectごとは，peopleごとに比べると若干率が悪いが，これはもともとのページが人間に関するページであり，projectクラスのインスタンスと他のクラスのインスタンスの間に直接的な連想構造が生成されなかったためである．

連想構造と利用者の知識を用いることにより，特別な自然言語処理を行わなくても，HTML文書からの概念抽出と整理がうまく行えることが分かった．

8. 議　論

本研究では，雑多で構造の不均質な情報源から情報を収集・整理するための手法を提案した．本手法における新規性は，データ構造としての連想構造にある．連想構造は，(1) 生データから容易に生成できることと，(2) 人間が直観的に理解しやすいことが特徴である．連想構造を用いることにより，多様な情報源からの情報の収集と整理を低コストで行える．

形態の雑多性に関しては，概念と外部参照データという2種類のユニットを導入することにより，電子化された多様な情報源の情報を統合的に扱うことができた．また，連想構造は記述が容易であるため，電子化されていない情報源からの情報獲得も容易であることが分かった．

表記の雑多性に関しては，辞書構造の導入と，知的情報統合機構のヒューリスティックを用いることにより，一部解消することができた．しかし，概念の多義性の問題や，日本語の表記のゆれについては本研究では扱わなかった．

構造の雑多性に関しては，WWW上のHTML文書のようなタグ付きの構造化テキストを例題として実験を行った結果，連想構造を用いることにより，HTML文書からの概念抽出と整理がうまく行えることが分かった．ここで実装した情報キャプチャ機構はキーワード

抽出を目的としたものであり，ほしい情報が文章で表されるような場合には異なるアルゴリズムを設計する必要がある．また，タグのついていないプレーンテキストに対しては今後の課題である．

9. 関連研究

本研究は，知識メディア[8]・オントロジー・エージェントによる仲介を基本技術として知識共有と再利用の基盤の確立を目的とする知識コミュニティプロジェクト[9]の1つである．本研究では，連想構造を基本構造とする知識メディアに焦点をあて，情報を収集，整理するシステムを開発した．

本研究の関連研究として，インターネット上の異質な情報からの情報獲得の研究[10)~14)]がある．また，WWW に関しては Yahoo をはじめとして，多くの検索エンジンが広く実用化されている．これらは主に情報収集の手法に焦点があてられているが，本研究のように収集した情報を統合・組織化するところまでは研究が進められていない．

本研究をネットワーク上の情報の抽出・加工の点から見ると，佐藤ら[15]の研究と関連する．この研究では，ネットニュースを対象として，抽出したい情報の雛型を設定し，スタイルなどを与えるアプローチがとられている．本研究では WWW の HTML 文書などを対象としている．本研究の情報キャプチャ機構は，利用者にドメイン知識がない場合や，最初にどんな情報がほしいか分からない場合も考慮して，不要な情報を取り込む可能性があるがどのようなドメインでも汎用的に取り込む段階と，ヒューリスティックを与えてより有用な情報を取り込む段階から構成されるアプローチをとった．

新聞記事からの情報抽出に関しては，MUC (Message Understanding Conference)[16),17)]において多くの実験評価が行われいる．MUC では開発者にあらかじめ見本データが通知されるため，情報抽出の精度が非常に高い．一方我々は，既存の多様な情報源の情報の整理を目的としているため，抽出精度は劣るかもしれないが，連想構造を用いて情報をゆるやかに取り込んだ後に，情報の構造化を図っている．

本研究を，人間の創造的問題解決プロセスを支援するコンピュータシステムという広義の発想支援システムの枠組み[18]でとらえると，KJ エディタ[19]や CAT1[20]などと類似性がある．これらは主にアイディアの発散・収束支援に焦点があてられている．一方，我々は既存の雑多な情報源の情報を整理する手法に焦点をあてている．

10. おわりに

既存の雑多で構造の不均質な情報源から情報を収集・整理する手法を提案した．基本となるアイデアとして雑多な情報をゆるやかに関連づける連想構造というデータ構造を用いた．この手法に基づき情報整理システム CM-2 を試作した．CM-2 では，(a) 既存の情報源から情報を取り込み，連想構造を生成する情報キャプチャ機構，(b) キーワードに基づき情報の切り出しと構造化を行う知的情報統合機構を実現した．CM-2 の有効性を実験によって確かめた．

今後は，異なる情報源から取得した大量の情報の統合（たとえば，人工知能に関して今回の WWW とネットニュース）などの課題に取り組む予定である．

参考文献

1) 国立国語研究所：現代表記のゆれ，国立国語研究所報告 75, 秀英出版 (1983).
2) Maeda, H., Koujitani, K. and Nishida, T.: A Knowledge Media Approach Using Associative Representation for Constructing Information Bases. *Proc. 9th International Conference on Industrial and Artificial Intelligence and Expert Systems (IEA/AIE-96)*, pp.117–126 (1996).
3) 糀谷和人：実データに基づくボトムアップなオントロジーの構築支援，奈良先端科学技術大学院大学修士論文 (1996).
4) 松本裕治ほか：日本語形態素解析システム JUMAN 使用説明書 version 2.0 (1994).
5) Brill, E.: Some Advance in Transformation-based Part of Speech Tagging, *Proc. 12th National Conference on Artificial Intelligence (AAAI-94)* (1994).
6) 前田晴美，西田豊明：知識メディアシステム CM-2 とそのユーザインタフェース，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.49–54 (1995).
7) マップルマガジン 64 奈良・大和路'94，昭文社 (1994).
8) Stefik, M.: The Next Knowledge Medium, *AI Magazine*, Vol.7, No.1, pp.34–46 (1986).
9) Nishida, T., Takeda, H.: Towards the Knowledgeable Community. *Proc. International Conference on Building and Sharing of Very Large-Scale Knowledge Bases 93*, Japan Information Processing Development Center, pp.157–166 (1993).
10) Levy, A.Y., Sagiv, Y. and Srivasava, D.: Towards Efficient Information Gathering Agents, *Working Notes of the AAAI Spring Symposium*

on Software Agents, pp.64–70 (1994).
11) Armstrong, R., Freitag, D., Joachims, T. and Mitchell, T.: A Learning Apprentice for the World Wide Web, *Working Notes of the AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, pp.6–12 (1995).
12) Balabanovi'c, M. and Shoham, Y.: Learning Information Retrieval Agents: Experiments with Automated Web Browsing, *Working Notes of the AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, pp.13–18 (1995).
13) Li, W.: Knowledge Gathering and Matching in Heterogeneous Databases, *Working Notes of the AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, pp.116–121 (1995).
14) Iwazume, M., Takeda, H., Nishida, T.: Ontology-based Approach to Information Gathering and Text Categorization, *Proc. International Symposium on Didital Libraries*, pp.186–193 (1995).
15) 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).
16) *Proc. 5th Message Understanding Conference (MUC-5)* (1993).
17) Grishman, R., and Sundheim, B.: Message Understanding Conference-6: A Brief History, *Proc.16th International Conference on Computational Linguistics (COLING-96)*, pp.466–471 (1996).
18) 國藤 進：発想支援システムの研究開発動向とその課題, 人工知能学会誌, Vol.8, No.5, pp.16–23 (1993).
19) 河合和久, 塩見彰睦, 竹田尚彦, 大岩 元：協調作業支援機能を持ったカード操作ツールKJエディタの評価実験, 人工知能学会誌, Vol.8, No.5, pp.583–592 (1993).
20) 角 康之, 堀 浩一, 大須賀節雄：テキストオブジェクトを空間配置することによる思考支援システム, 人工知能学会誌, Vol.9, No.1, pp.139–147 (1994).

(平成8年3月11日受付)
(平成9年1月10日採録)

前田 晴美（学生会員）

1986年京都大学文学部哲学科心理学専攻卒業. 同年富士通（株）入社. 1993年英国マンチェスター科学技術大学計算機学科修士課程修了. 現在奈良先端科学技術大学院大学情報科学研究科博士後期課程2年. 知識メディア, 知識共有の研究に従事. 人工知能学会, 日本認知科学会, 計測自動制御学会ヒューマン・インタフェース部会, AAAI各会員.

糀谷 和人

1994年大阪大学工学部電子工学科卒業. 1996年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了. 同年オムロン（株）入社. 新事業開発センタファジィ推進室所属. 知識獲得, データマイニングの研究に従事. 人工知能学会会員.

西田 豊明（正会員）

1977年京都大学工学部情報工学科卒業. 1979年同大学院修士課程修了. 京都大学工学部助手, 助教授を経て, 1993年奈良先端科学技術大学院大学情報科学研究科教授, 現在に至る. 京都大学工学博士. 1984年から1年間Yale大学客員研究員. 1995年から科学技術庁金属材料技術研究所客員研究官. 知識の共有と再利用, 知識メディア, 定性推論の研究に従事. 1988, 89, 95年人工知能学会全国大会優秀論文賞受賞. 1988年度人工知能学会論文賞受賞. 1990年本学会創立30周年記念論文賞受賞. 著書「自然言語処理入門」（オーム社),「定性推論の諸相」（朝倉書店）など. 人工知能学会など各会員. IJCAI-97 Video Track Chair, 京都21委員など.