概念の連想表現に基づく情報空間形成支援

奈良先端科学技術大学院大学情報科学研究科
○前田晴美、糀谷和人、西田豊明

Creating Information Space Based on an Associative Indexing Method

○ Harumi MAEDA, Kazuto KOUJITANI and Toyoaki NISHIDA
Graduate School of Information Science, Nara Institute of Science and Technology

Abstract: There exist various kinds of information sources such as WWW pages, personal memoranda, hypertexts and image files. In this paper, we present a knowledge media approach using *associative indexing* as a framework of information representation to integrate a wide variety of heterogeneous information gathered from multiple information sources into personal information space.

We then present a knowledge media information base system called CM-2 which provides users with a means of accumulating, sharing, exploring and refining conceptually promiscuous information gathered from vast information sources. We describe the system's three major facilities; (a) an *information capture facility* which allows users to gather information from various kinds of information sources and create CM-2 information bases, (b) an *associative retrieval facility* which gives users access to multimedia information stored in the information base through associative indexing mechanisms, and an (c) an *information integration facility* which helps users integrate multiple information bases into personal information space from the user's point of view.

We discuss the strength and weakness of our approach by analyzing results of experiments.

1 はじめに

研究活動に代表される人間の創造的思考活動には、さまざまな知識や情報が必要である。例えば、自分が過去に作成した論文やOHP、アイディアメモや、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典や辞書、オンラインデータベースやWWW(World Wide Web)など、種類も形態も異なる雑多な情報である。文書化されていない頭の中にあるアイディアの断片や常識なども非常に重要である。

創造的な思考活動に集中するために、世の中に存在するさまざまな種類の知識情報を簡単に収集・整理でき、それらを論文作成などの知的作業に利用できるようなシステムの実現が望まれる。

しかし、実世界に既に存在している膨大な規模の異なる種類の情報を扱うためには、既存の関係モデルやオブジェクトモデル、一階述語論理などを用いて一から情報ベースを形成することは人間の負荷が高い。また、WWW に代表されるようにマルチメディア情報をハイパーテキスト表現で扱う方法は、人間にとってわかりやすいがコンピュータが情報を加工することが難しい。

そこで我々は、さまざまな種類の知識や情報を統合して扱うための、人間向きメディアとコンピュータ向きメディアの中間表現として、連想表現を基本構造とする知識メディア [1] を提案する。連想表現は概念間を連想関係で結ぶ情報表現の方法であり、従来の知識情報表現と比較するとはるかに単純で人間向きであり、コンピュータにも扱いやすい。

我々は本アプローチに基づき知識メディア情報ベースシステム **CM-2** を試作した。CM-2 には従来のシステムにはない新しいしくみが必要であり、以下の機構を開発した。

- 自然界に存在する内容も構造も違うさまざまな情報を取り込むための情報キャプチャ機構
- 情報ベースのマルチメディア情報を連想的に検索するための連想検索機構
- 情報を個人の視点から整理・利用するための知的情報統合機構

以下では、2 節で CM-2 と連想表現の概要を紹介し、3、4、5 節で CM-2 の 3 つの手法について説明し、以下は実験と議論を述べる。

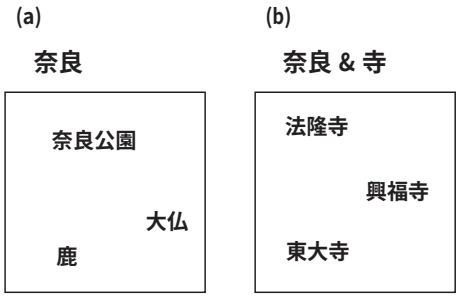


図 1: CM-2 における連想表現

2 CM-2 における連想表現

2.1 連想表現

連想表現は、概念間の連想による連結を基本構造とする弱構造の知識メディアである。知識メディアの最小単位をユニットと呼ぶ。以下に主要なユニットの種類を示す。

- 概念：概念を示すユニット。
- 外部参照データ：外部データベースのマルチメディア情報を参照することを示すユニット。
- 連想：概念または外部参照データ間を連想関係で連結するユニット。

１つ以上の概念または外部参照データから１つ以上の概念または外部参照データが連想される。連想元を key、連想先を value と呼ぶ。本稿では CM-2 における論理的な連想表現を図 1(a),(b) のように記述する。図 1(a) では key 「奈良」から value 「奈良公園」「鹿」「大仏」が連想されている。図 1は key 「奈良」と「寺」から value 「東大寺」「興福寺」「法隆寺」が連想されている。

連想には単純な連想関係に加えて、特別な形態として、クラス・インスタンス関係、同義語関係を定義することもできる。

2.2 CM-2 の概要

CM-2 [1]は、膨大な情報源から集められたさまざまな異質な情報を蓄積、共有、探訪するための、個人やグループのための情報ベースシステムである。

CM-2 の情報ベースは、内部知識ベース (概念ベース) と外部データベース (マルチメディアデータベース) から構成される。利用者はワークスペース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。図 2に CM-2 のアーキテクチャを示す。

CM-2 のユーザインタフェースはカード型のワークスペースとエージェントから構成される。利用者は、自由にメモを書いたり考えをまとめたりする感覚で、ワークスペース上で思考や情報収集などのさまざまな情報活動を行う。ワークスペースには自律的なソフトウエアモジュールであるエージェントが住んでおり、さまざまな情報サービスを提供する。

ワークスペースのデザインは概念マップに基づいており、概念間の関係は有向矢印として表示される。概念の横にボックスが表示されている時は連想される概念が存在することを示し、ボックスをマウスで選択することにより表示することができる。(近傍検索; 4.1節参照)

また、ワークスペースを通してマルチメディア情報へアクセスできる。外部参照データの横に表示されるボックスをマウスで選択すると、データに応じてワークスペースに表示または、適切な外部ビューワが活性化されて表示される。

2.3 実世界における概念の乱雑さ

現実の世界においては、概念にどのような名前付けを行うかは個人によって違う。同じ概念に違う名前をつけることもあるし、違う概念に同じ名前をつけることもある。また、知識や情報の構造は個々によって違う。我々はこのような現象を概念の乱雑さと呼ぶ。

[1] “CM” は 我々の長期的な研究目標である “Contextual Media(文脈メディア)” の意味である。

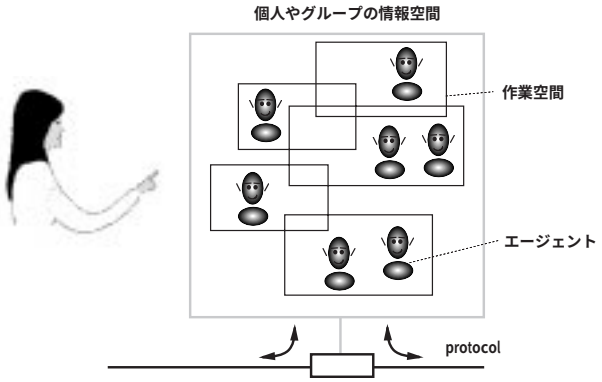


図 2: CM-2 のアーキテクチャ

我々は、概念の乱雑さを扱う指針として、用語の統制化を行ったり厳格なスキーマの設計を行うよりもむしろ、概念の名前付けや何から何が連想されるかは、情報ベース作成者の自由とする。このような「寛大さ」は、情報ベース構築における時間と労力を節約するために有効であると考える。連想表現は極めて単純であり、実世界の生データから情報ベースを作成することは非常に容易である。

しかし、情報ベースの構造設計を作成者に委ねることは、利用者側から見ると他人が作成した情報ベースの構造が理解しにくいため検索が不便である。

次節以降では、 CM-2 の 3 つの主要な機構について述べる。各機構は概念の乱雑さを扱うためのヒューリスティックを含んでいる。また、各機構は機能のまとまり毎にエージェントとして実装されている。

3 情報キャプチャ機構

利用したい情報は、書籍や個人的な記憶、 WWW などさまざまなところに存在する。自然界に存在する内容も構造も違う情報源から情報を取り込み、情報ベースを生成するための機構である。

自然言語からの情報取り込みには膨大なヒューリスティックが必要かもしれない。我々は WWW に焦点をあて、情報キャプチャの手法を検討した。

WWW の記述言語である HTML はかなり柔軟な言語であり、作成者に厳格な情報構造化を要求しない。例えば <h1> などの見出し表現や <dl> や <ul> などの箇条書表現を使って文書を構造化してもよいししなくてもよい。また、見出し表現は、見出しとしてではなく、単にブラウザ表示時の文字の大きさを変えたいために使用されることもある。さらに、作成者と利用者によって重要な情報が違うこともある。

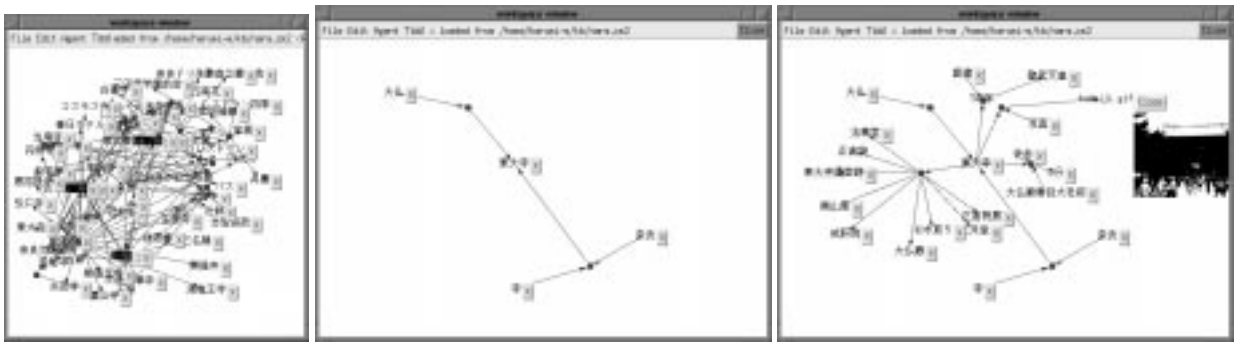
HTML のタグ付けに着目したヒューリスティックを用いて予備実験 (6節参照) を行ったところ、タグだけでは不十分なことや、クラス・インスタンス関係も含めて必要な知識を獲得するためには、情報抽出のためのドメイン知識や自然言語処理が必要であることなどがわかった。しかし、利用目的に応じてその都度固有のプログラムを開発することは手間がかかる。そこで、ある程度汎用的なキャプチャ機構を開発して一旦情報ベースを生成してから、情報編集機構 [2] や、知的情報統合機構 (5節参照) を用いて、利用者にとって使いやすい情報ベースを構築する方針とした。

現在までに、情報キャプチャ機構として、実験に記述した WWW からの取り込みの他に、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、日本経済新聞全文記事データベース CD-ROM、 Lisp プログラム、 UNIX ファイルシステムなどから CM-2 情報ベース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頭の中にある知識などから CM-2 情報ベースを生成するため、ワークスペース上から概念と連想を入力できるようにした。

4 連想検索機構

CM-2 の情報検索には、キーワード検索と連想検索がある。本節では、情報ベースから連想的にマルチメディア情報を検索する連想検索機構について述べる。



(a) 近傍検索 (問題のある場合)　　(b) 活性境界探索型連想検索　　(c) 近傍検索

図 3: 連想検索機構の画面例

4.1 近傍検索

近傍検索は画面上から概念を選択することにより連想で結ばれた概念を検索するものである。例えば、図 1(b) の連想において利用者が「奈良」を選択すると、「寺」「東大寺」「法隆寺」「興福寺」が表示される。

近傍検索では、選択された概念に非常に多くの key や value が連結されている時に表示に時間がかかり、概念が見えにくくなるという問題が生じる。図 3(a) は研究室で手作業で構築した奈良観光情報ベースにおいて、「奈良」を選択した場合の画面例である。このような問題を解決するのが次で述べる活性境界探索型連想検索である。

4.2 活性境界探索型連想検索

活性境界探索型連想検索は、情報ベースの情報をダイナミックに検索する。複数の概念を選択すると、概念の周囲に活性化の輪が広がり、輪が接触したところの概念を検索する。活性化の距離は 1 つの連想間を 1 としている。

例えば、「奈良で大仏のある寺は？」という質問文を入力すると、「奈良」「大仏」「寺」から連想をたどることにより解「東大寺」を検索し、図 3(b) のような結果をワークスペースに表示する。

以下にアルゴリズムの概要を示す。

- ステップ 1: 質問文を解析して概念を抽出する。

  (例: 質問文「奈良で大仏がある寺は？」より「奈良」「大仏」「寺」が抽出される。)

- ステップ 2: 抽出された概念から等距離で連想をたどり、積集合を解とする。距離は key-value 間を距離 1 とし、距離 1,2 まで計算する。

  (例: 「東大寺」が見つかる。)

- ステップ 3: 抽出された概念、解、これらにリンクされる連想をワークスペースに表示する。

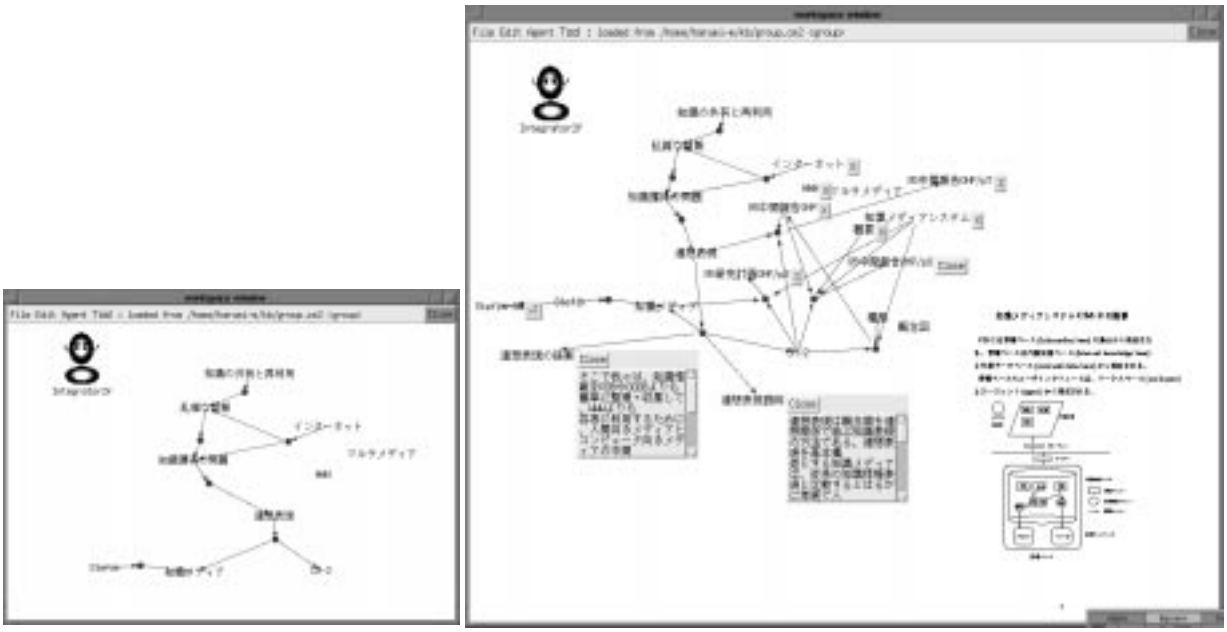
  (例: 「奈良」「大仏」「寺」「東大寺」が表示される。)

また、図 3(b) で「東大寺」を選択することにより、近傍検索により図 3(c) のように「東大寺」から連想されるさらに詳しい情報を得ることができる。このように近傍検索と活性境界探索型連想検索は相補的な連想検索機構であり、二つの組合せにより情報ベースの内容の効率的探訪が行える。

5 知的情報統合機構

知的情報統合機構は、複数の情報源から生成された情報ベースを統合して、情報を個人の視点から整理、利用する機構である。

例えば、創造的思考活動においては、アイディアを出したり考えをまとめる途中で関連文献を調べたくなったり、自分やグループのメンバが作成した論文や OHP が見たくなることがある。このような時に、自分が過去に作成した情報ベースや他人が作成した情報ベースの情報を自分の情報空間にとりこんで表示できればたいへん便利である。しかし、他の情報ベース、特に他人の作成した情報ベースは構造がよくわからないために利用が難しい。



(a) 統合前の情報空間　　(b) 統合後の情報空間

図 4: 知的情報統合機構の画面例 1

知的情報統合機構は人間のこのような個人やグループの情報活動を支援する。図 4(a) には第一著者が本論文執筆にあたり作成した思考空間を示す。図 4(b) は、図 4(a) の空間を視点として利用し、第一著者が過去に作成した論文情報ベースとOHP 情報ベース、それから第三著者の作成したサーベイメモ情報ベースを統合して、近傍検索を行っている様子である。

別の例として、自分の興味のある人工知能関連のプロジェクトが知りたいが、そのようなデータベースはどこにも存在しないとする。このような場合人間は、個人の知識やその他の情報を参照しながら、大学の人工知能関連の WWW ページにアクセスするかもしれない。しかし、自分の知識の構造や各 WWW ページの構造は異なるため、自分のほしいかたちで情報を整理することは容易ではない。本機構では、利用者は個人の知識や本で得た情報、 WWW ページなどを統合し、知りたい項目を入力すると、動的に情報の整理統合が行われる。統合結果はワークスペースだけでなく HTML ファイルなどの別の形式に変換して外部ビューワに表示することもできる。(図 5参照。)

以下にアルゴリズムの概要を示す。ステップ 1 は、図 4及び図 5の例の共通アルゴリズムである。ステップ 2 以降は図 5の例のみで、項目を入力して情報を統合表示する時の処理で、情報ベースにクラス・インスタンス関係の連想が含まれていることを前提としている。

- ステップ 1: 視点となる情報ベースに基づき他の情報ベースを統合する。この時、視点となる情報ベースに関連するユニットだけ統合 (抽出統合) または全てのユニットの統合 (全体統合) の選択が可能である。また、元の視点と適合しない連想ユニットは動的にくみかえを行う。以下に主要なヒューリスティックを示す。
  - 同名概念統合: 同名の概念を統合する。
  - 辞書参照概念統合: 同義語関係を用いて記述した辞書を参照して概念を統合する。
  - 包含的連想生成: ある概念の名前が他の概念の名前に含まれている時、その概念を key として他の概念を value とする連想を生成する。
    (例: 利用者の興味 (knowledge representation, agent, integration) が記述されている視点に基づき、他の情報ベースから抽出できる情報のみを統合する。)
- ステップ 2: 連想をたどることにより解を探し、新しい連想を生成する。
  1. 最初の key となる概念を、情報ベース中の入力項目またはその同義語のクラスのインスタンスの中から見つける。
     (例: 「プロジェクト」クラスのインスタンスである「Knowledge Sharing Technology」「Adaptive Intelligent Systems」「How Things Work」」を解とする。)

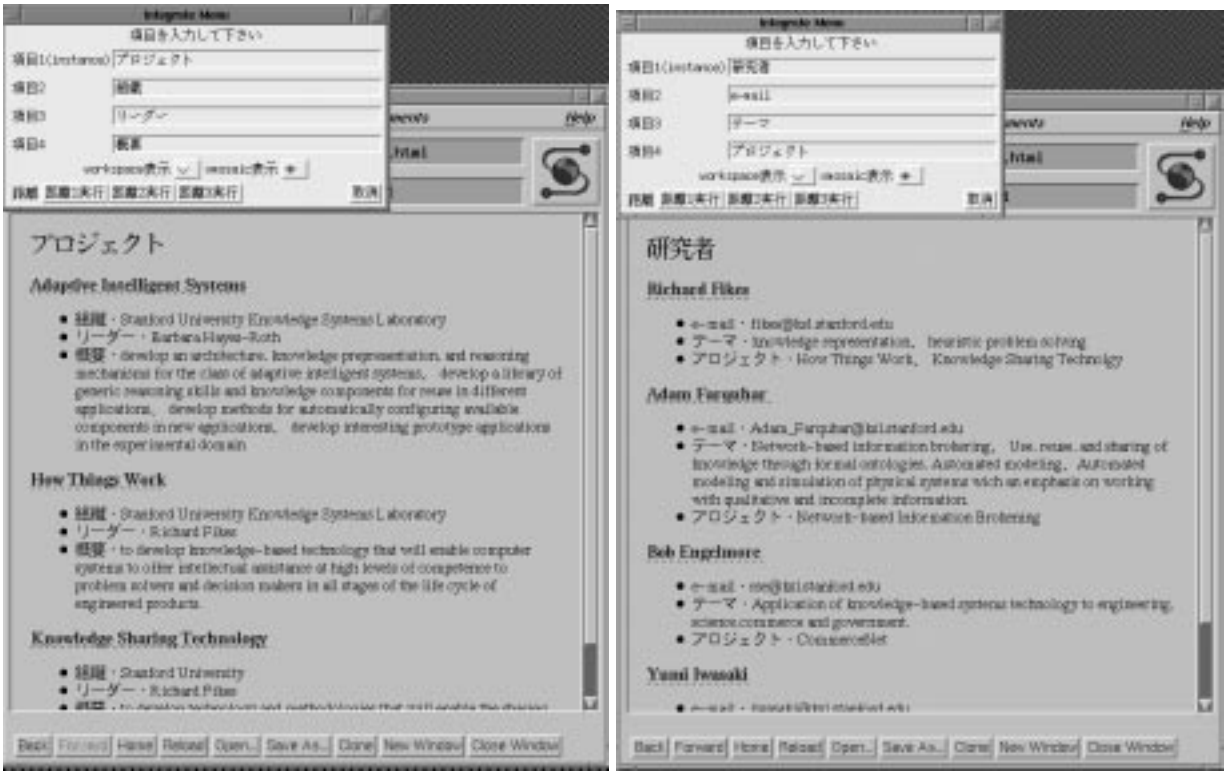


図 5: 知的情報統合機構の画面例 2

2. 2 つ目以降の key となる概念は、入力項目またはその同義語のクラスのインスタンスと、最初に特定された概念から連想をたどることにより探す。

   (例: 「Knowledge Sharing Technology」の「リーダー」に関しては、「Knowledge Sharing Technology」の近傍にある「研究者」クラスのインスタンスである「Richard Fikes」を解とする。)

3. 新しい連想を生成する。

   (例: 「Knowledge Sharing Technology &リーダー→ Richard Fikes」を生成する。)

- ステップ 3: 生成された連想をワークスペースに表示または HTML ファイルを生成し、 WWW ブラウザを起動表示する。

6 実験

6.1 情報キャプチャ機構

6.1.1 タグに着目した情報取り込み予備実験

我々は見出し表現、段落表現、箇条書表現、アンカー表現などのタグに着目したヒューリスティックを用いて予備実験を行った。奈良先端大のホームページからリンクのはられている文書のうち 11HTML 文書 (日本語) を取得して 342 個のユニットを生成した。その中で 295 個 (約 86%) が適切であると評価した。

この時以下の課題が残った。

1. 見出しや箇条書からそのまま概念を抽出すると、「バイオサイエンス研究科及び遺伝子研究センターホームページ」などの長い情報を概念として抽出してしまう。このような冗長な概念は情報統合のためには不適切である。

2. 何もタグ付けされていない文に重要な概念が含まれていても抽出することができない。

3. 「Click here for the home page in English.」ような WWW に特徴的なハイパーリンク表現の抽出ように不要な概念を抽出する。

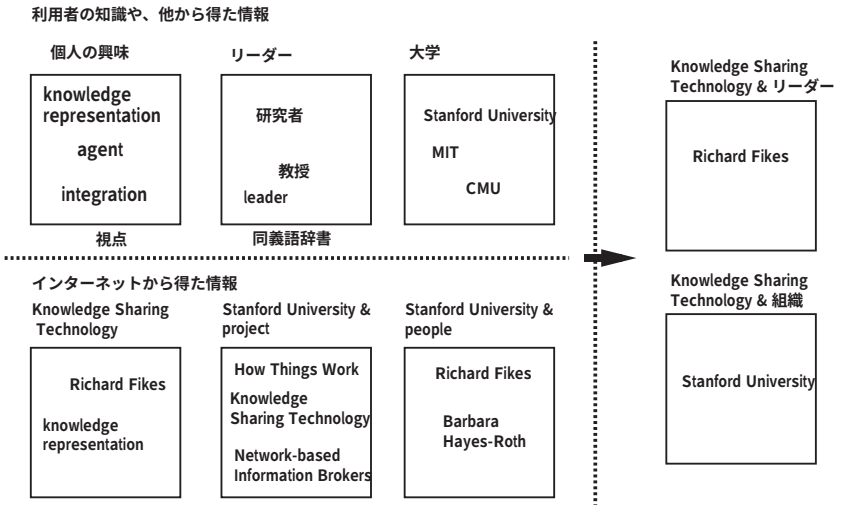


図 6: 知的情報統合機構のアルゴリズム概念図

4. クラス・インスタンス関係や同義語関係を抽出するためには、ドメイン知識の利用や自然言語処理が必要である。

6.1.2 形態素解析を利用した情報取り込み実験

上記 1,2,3 の問題を克服するために、形態素解析を利用して、タグだけでなく文書全体の名詞及び名詞句の抽出を行った。スタンフォード大学知識システム研究所のホームページからリンクのはられたページの中の 8 つの HTML 文書 (英語) について 537 個のユニットを生成した。その内概念 450 個について評価したところ、 435 個 (97%) が適切であった。不適切な概念は、名詞句の生成失敗によるものである。この問題はどの品詞をつなぎあわせて名詞句とするかを調節することによってほぼ解消できると予測される。以上の結果は、意味理解を伴わない HTML 文書からの網羅的なキーワード抽出と連想付けに関しては自動化が容易であることを示している。

今後は概念だけでなく、外部参照データの抽出及びクラス・インスタンス関係も含めた概念の関係をどこまで汎用的に抽出できるかを検討する。また、ドメイン知識や自然言語処理を利用して個別に必要な知識を抽出するプログラムを開発する方法との比較検討も行う。

6.2 連想検索機構

手作業で構築した奈良の観光案内情報ベース (概念 1286 個, 連想 862 個) に対して、活性境界探索型連想検索の実験を行った。事前の予備実験 (10 問) により、最短距離で正答が得られた場合距離をのばしても正解がふえないことと、距離 3 では正答がほとんど得られない (正解以外の解がほぼ必ずまざる) ことがわかったため、最初は距離 1 で探索を行い、正答が得られない場合のみ距離 2 探索を行った。質問に対する回答を以下の基準で主観的に評価した。

- 得られたすべての解が正解である場合：正答
- 解なしあるいは解に正解が含まれない場合：誤答
- 解に正解及び正解以外の解が含まれる場合：その他
  その他は、正解以外に重大な間違いを含む場合 (その他 2) と含まない場合 (その他 1) に分類した。

30 問の質問のうち、距離 1 探索での正答は 16/30(約 53%) である。距離 1 と距離 2 の総合評価では正答は 17/30(約 57%) である。正答にその他 1 を含めると、正答率は約 80% となり、結果は良好であると評価する。また、正解以外の場合でも解にいたる連想が表示されるので、利用者はその解が正しいかどうかを理解することができる。

6.3 知的情報統合機構

6.3.1 グループの情報ベースの統合実験

5節の図 4の例において、同義語辞書を用いずに予備的な全体統合実験を行った。統合元の情報ベースの概念数が 10 個であるのに対して、統合する 3 つの情報ベースの概念の総数が 129 個であった。結果は、連想検索により特に不便さを感じる

ことなく統合された情報にアクセスすることができた。統合の際、同名概念統合が 9 回、包含的連想生成が 63 回発生した。また、統合したい概念の中で、現状のヒューリスティックで統合できなかったのは「cm2」と「CM-2」のように表記にゆれがある 3 個の概念のみであった。

以上の結果より、包含的連想生成が強力なヒューリスティックであることが示唆される。もっとも、この実験で統合した情報ベースは内容の関連性がかなり高いことと、作成者が同じ研究グループであるために概念の乱雑さが少ないことが統合がうまくいった大きな理由であると考えられる。

6.3.2 WWW の情報と個人の知識の統合実験

「人工知能、知識表現」などのキーワードを視点として、奈良先端大情報科学研究科教官紹介の WWW ページを情報ベース化したもの (クラス・インスタンス関係は手作業で与えてある) を抽出統合すると、与えたキーワードに関連した教官名とテーマを抽出・表示が行えた。

今後は、情報キャプチャ機構とあわせて総合的に評価を行い、他人が作成した情報の統合をどの程度自動化できるか試みる。

7 議論

我々は、個人やグループの創造的思考活動を支援するために、世の中に存在する複数の情報源から情報を獲得、整理、統合して個人的な情報空間を形成する手法に焦点をあてた。

本研究の背景には、コンピュータが扱いやすい情報表現と人間が利用しやすい情報表現のトレードオフがある。厳格な情報モデルを用いるほどコンピュータにとって扱いやすいが、人間にとって情報表現の利用が難しくなる。反面人間にわかりやすいことを追求しすぎるとコンピュータが知的な処理を行いにくくなる。

我々は、コンピュータにも人間にも利用しやすい中間表現としての連想表現を提案し、本アプローチに基づき、実際に動作するシステムを開発し、実験を行った。

情報キャプチャに関する実験では、既存の情報を概念レベルで抽出して連想関係をつけることが非常に容易であることがわかった。連想検索に関する実験は、構造のわからない情報ベースからでも連想的に検索が行えることを示した。解が正解でなくても解にいたる連想が表示されるので人間は理解できることも示した。情報統合に関する実験は、他人が作成した情報を個人の視点から利用できることと、情報を他の形式の情報にも簡単に変換できることを示し、さらに情報を統合することで新しい価値をもった情報を作成できる可能性も示した。

いくつかの課題も明らかになった。検索や統合のヒューリスティックを補い、コンピュータに知的な処理をさせるためには、連想以外の関係 (クラス・インスタンス関係や同義語関係など) が必要であるが、自然情報源から汎用的に意味情報を抽出するのは容易ではない。また、概念や関係の数が増えると概念マップは操作性や可視性が悪くなるので工夫が必要である。

8 まとめ

創造的思考活動に必要なさまざまな自然情報源の情報を扱うために、人間向きメディアとコンピュータ向きメディアの中間表現として、連想表現を基本構造とする知識メディアを提案した。我々のアプローチは世の中の異質な情報を労力と時間をかけずに収集、統合整理し、個人やグループの情報空間を形成するための枠組を提供した。

本アプローチに基づき知識メディア情報ベースシステム CM-2 を試作した。 CM-2 の主要な特徴は、 (a) 自然界に存在するさまざまな情報から情報ベースを生成するための情報キャプチャ機構、 (b) 情報ベースのマルチメディア情報を連想的に検索するための連想検索機構、 (c) 個人の視点から複数の情報ベースの情報を統合・利用するための知的情報統合機構である。

CM-2 は Common Lisp で記述され、 GUI は Gcl/Tk を使用しており、現在 UNIX/X 環境で動作している。

参考文献

[1] Mark Stefik. 1986. The next knowledge medium. *AI Magazine* 7(1):34-46.

[2] 糀谷和人、前田晴美、西田豊明. 1995. 弱構造知識メディアを用いた情報ベース構築支援. 信学技法, Vol.95.No.265.pp.63-70.