連想構造を用いた情報収集・整理支援

Gathering and Reorganization of Information Using Associative Structures

前田 晴美、 糀谷 和人、西田 豊明

Harumi Maeda, Kazuto Koujitani and Toyoaki Nishida

奈良先端科学技術大学院大学 情報科学研究科

Graduate School of Information Science, Nara Institute of Science and Technology

In this paper, we present a method to reorganize diverse information obtained from heterogeneous information sources such as WWW pages. The basic idea is the use of a plain information representation called *associative structures* which enable to link information chunks using associative indexing. We have developed a system called CM-2 which helps users to gather and reorganize information from existing information sources. We describe the system's two major facilities; (a) an *information capture facility* which allows users to gather information from heterogeneous information sources and generate CM-2 associative structures and (b) an *information integration facility* which helps users to reorganize information from the user's point of view We verify our approach by analyzing results of experiments.

1. はじめに

世の中には情報が溢れている。特にWWW(World Wide Web) 上では爆発的な勢いで情報が増加している。インターネット上の情報源から情報を収集するシステムは数多くあるが、収集した情報を個人の視点から整理できるシステムはこれまでにほとんどなかった。

本論文では、 WWW のページのような雑多で構造の不均質な情報源から情報を収集・整理する一つの手法を提案する。基本となるアイディアは構造の異なる情報をゆるやかに関連づけるための連想構造とよぶデータ構造を用いることである。連想構造を用いることにより、生データから情報を抽出し、個人の視点から整理することが簡単に行える。

我々は，連想構造を用いて情報を収集し，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、 WWW の URL を入力として、個人の視点から情報を切り出し・構造化して提示することができる。

図1に、 CM-2 による情報収集・整理の概要を示す。図1の WWW ページ例は、人工知能研究者 2 人のホームページの一部 [1], [2] である。直感的に明らかなように両者のページの情報の構造や表現方法は異なっている。 CM-2 では、これらのページから情報を抽出しての連想構造を生成し、得られた連想構造をヒューリスティックスを用いて統合する。利用者が「研究者」や「プロジェクト」などの視点を入力することにより、研究者毎やプロジェクト毎に情報を整理して提示を行なう。

本論文の以下の構成は次のとおり。 2 節では連想構造について述べる。 3、 4 節では情報収集・整理を支援する CM-2 の 2 つの機構について説明する。 5 節では実験結果を示し、有効性について議論する。 6 節では関連研究と比較する。

2. 連想構造

CM-2 では情報の基本構成要素を**ユニット**と呼ぶ。

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付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。

これらのユニットは、連想構造によって関連付けられる。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 の間の連想関係は精密に定義されたのではなく、多分に主観的なものであることを許している。これは既存の情報の雑多性、多様性に対応することを狙ったものである。

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。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、辞書構造を定義することもできる。 IS-A 構造 (図 2(c)) においては、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 構造は推論に使用される。辞書構造は、利用者の語彙を定義するもので、概念ユニット間の変換辞書として使用される。例えば、図 2(d) の辞書構造がある場合、利用者が「奈良先端大」と入力すると「NAIST」というユニットに変換が可能である.

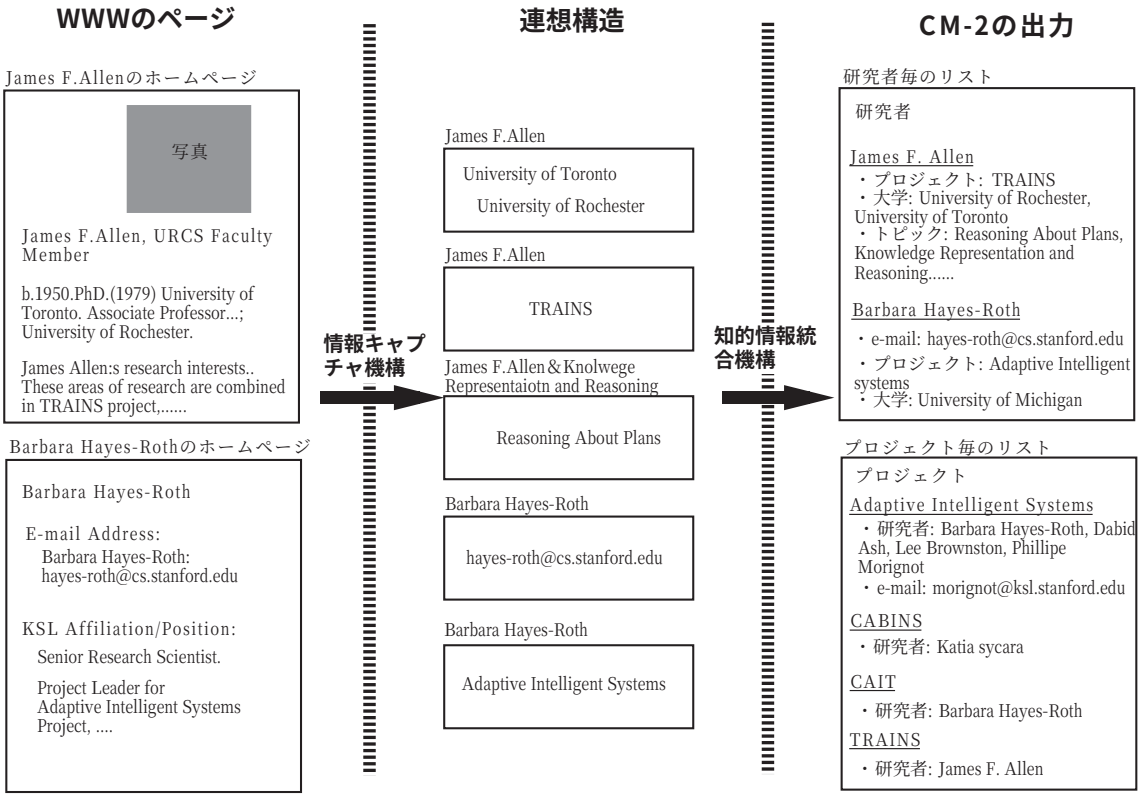

図 1: CM-2 による情報収集・整理の概要

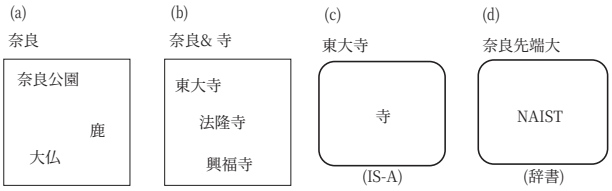

図 2: 連想構造

3. 情報キャプチャ機構

情報キャプチャ機構は、既存の雑多で構造の不均質な情報源から情報を取り込み、CM-2 の連想構造を生成する.

WWW で使用される HTML は作者に情報の構造化を要求しないため、HTML 文書から利用者にとって有用な情報を抽出する方法は自明ではない。我々は、形態素解析と HTML 構造解析を用いて名詞句主体のユニットと連想構造を生成してから、ドメイン知識を利用して重要なキーワードを抽出する方針とした。図 3に情報キャプチャ機構のアルゴリズムを示す。

step 3 で用いられるドメイン知識は、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簡単なヒューリスティックスを中心とする。

3.1 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が行われるのか説明する (図 4)。

step 1 URL を解析して HTML 文書を取得する
step 2 形態素解析と HTML 構造解析を用いてユニットと連想構造を生成する
  step 2.1 形態素解析 (日本語:juman、英語:Brill's Rule Based Tagger[3]) に基づく簡単な概念抽出規則を用いて名詞句とそれに準拠するものを抽出し、ユニットを生成する
  step 2.2 HTML の構造を解析して連想構造を生成する
step 3 ドメイン知識を用いて IS-A 構造を生成し、ユニットと連想構造を修正する
  step 3.1 ユニットからドメイン知識を用いて IS-A 構造を抽出・生成する
  step 3.2 step 2 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

図 3: 情報キャプチャ機構のアルゴリズム

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」、「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.

「James F.Allen」は「James」という英語名の単語

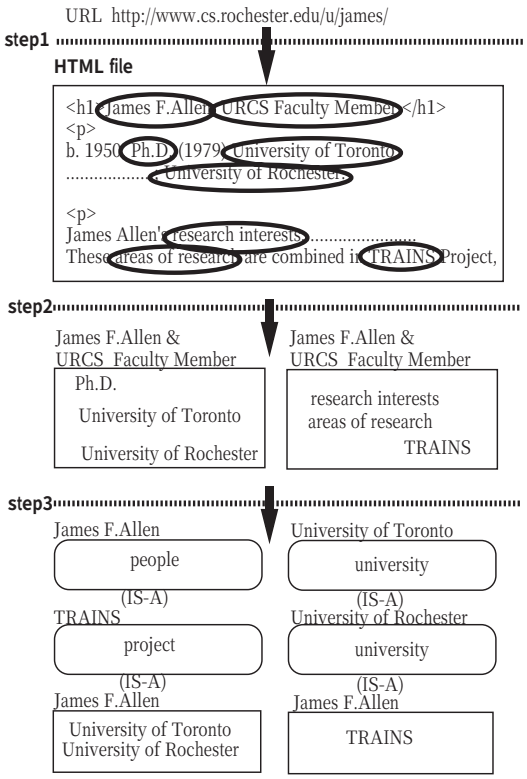

図 4: 情報キャプチャ機構アルゴリズム動作例

を含むため「people」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」であると推論し、IS-A 構造を生成する。

生成された連想構造の中には「URCS Faculty Member」「Ph.D.」に関連する IS-A 構造がないため、これらを削除する。また「James F.Allen」と「URCS Faculty Member」を key とする連想構造からこれらのユニットを削除・修正する。

4. 知的情報統合機構

知的情報統合機構は、与えられた視点から情報の切り出しと構造化を行う。基本的なアルゴリズムは、ヒューリスティックスを用いて CM-2 のユニットと連想構造を統合する **step1** と、パス探索を用いて利用者の入力に応じて新しい連想構造を生成して結果を表示する **step2** から構成される。

step 1 で用いられるヒューリスティックスの中、主要なものを表 1に述べる。

4.1 例

利用者が図 1のようなプロジェクトリストを出力したいとする。入力項目 1「プロジェクト」、入力項目 2「研究者」を入力した場合の動作例を示す (図 5参照)。

「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の中から、「TRAINS」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、「TRAINS」と連想構造でつながりがあるユニット「James F.Allen」を value として、連想構造を生成する。このように生成された連想構造をリストとして表示する。

表 1: 知的情報統合機構の主要ヒューリスティックス

同名概念統合	同名の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が他の概念ユニットの名前に含まれている時、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 が同じ時、連想構造を統合する

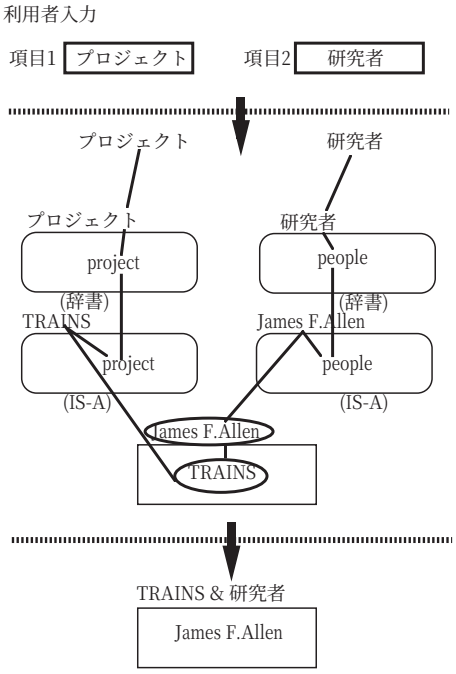

図 5: 知的情報統合機構 (step2) アルゴリズム動作例

5. 実験と考察

情報キャプチャ機構と知的情報統合機構の統合的な実験を行なった。

人工知能研究者のホームページ 100URL を入力して、利用者が項目を入力すると、図 1のような研究者毎リスト、プロジェクト毎リストを自動的に出力した。生成されたユニットは 764 個、連想構造は 625 個、IS-A 構造は 755 個である。情報キャプチャのために与えた知識は、people、project、e-mail、university、department、laboratory、topic、など

表 2: 実験結果

		適合率	再現率
テスト1	people	93%	85%
	people以外のクラス	89%	83%
テスト2	project	78%	82%
	project以外のクラス	65%	69%

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

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

7 つのクラスに関して 288 個である。

people 毎に残りのクラスを統合するテスト 1、project 毎に残りのクラスを統合するテスト 2 の結果を表 2に示す。

この結果は、形態素解析と HTML 構造解析及び簡単なドメイン知識を用いることにより、特別な自然言語処理を行なわなくても、 HTML 文書からの重要なキーワード抽出と視点変換がほぼ満足できる確度で行えることを示している。

project 毎は people 毎に比べると若干率が悪いが、これはもともとのページが people のページであり、project と他のクラスの間の直接的な連想関係がないためである。

この実験における誤りは以下のように分類される。

- **step 2** における名詞句生成失敗によるユニットの生成の誤り (例: プロジェクトである概念「How Things Work」は、「How」が副詞であるため名詞句とみなされず、ユニットが生成されなかった)
- **step 3** における HTML 構造解析失敗による連想構造生成の誤り
- **step 3** におけるドメイン知識不足による IS-A 構造抽出の誤り

今後の課題を以下に述べる。(1)**step 2** において、名詞句生成のためのヒューリスティックスを導入する。(2) HTML 構造解析の精度を向上させる。(3) キャプチャ中に得た知識を IS-A 構造として情報キャプチャ機構に再利用できるしくみをつくる。(4) 実験量をふやして検証する (クラス、 URL、他ドメインなど)。

6. 関連研究

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 [4],[5],[6], [7],[8] がある。また、 WWW に関しては Yahoo をはじめとして、多くのサーチ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個人の視点から統合整理するところまでは研究が進められていない。

本研究は情報抽出の点から見ると、佐藤ら [9] の研究と関連する。この研究においては、あらかじめ抽出したい情報ターゲットを設定し、構造情報や必要な知識を与えて精度をあげるアプローチがとられている。本研究では、利用者がドメイン知識がない場合や最初からほしい情報がわからない場合も考慮して、不要キーワード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み可能な段階 (step2) と、ドメイン知識を与えてより有用な情報を取り込む段階 (step3) から構成されるアプローチをとった。

7. おわりに

本論文では WWW に焦点を当て、既存の雑多で構造の不均質な情報を収集・整理する手法を提案した。基本となるアイディアは雑多な情報をゆるやかに関連づけるための**連想構造**とよぶデータ構造を用いることである。連想構造を用いることにより、既存の情報源から情報を抽出し、個人の視点から整理することが簡単に行なえる。我々は、連想構造を用いて情報を収集し、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、 URL を入力として、個人の視点から情報を切り出し・構造化して提示することができる。 CM-2 の有効性を実験によって確かめた。

参考文献

[1] http://www.cs.rochester.edu/u/james/

[2] http://www-ksl.stanford.edu/people/bhr/index.html

[3] Brill, E.: Some Advance in Transformation-Based Part of Speech Tagging, *Proceedings of the Twelefth National Conference on Artificial Intelligence (AAAI-94)*, (1994).

[4] 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).

[5] 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).

[6] 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).

[7] 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).

[8] Iwazume, M., Takeda, H., Nishida,T.: Ontology-based approach to information gathering and text categorization, *Proceedings of International Symposium on Didital Libraries*, pp.186–193 (1995).

[9] 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).