文脈メディアによる弱構造情報の提示と構造化

糀谷 和人・前田 晴美・西田 豊明
(奈良先端科学技術大学院大学 情報科学研究科)

Representation and Refinement of Information Base with a Contextual Medium

Kazuto Koujitani, Harumi MAEDA, Toyoaki NISHIDA
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-01 Japan

The contextual medium CM-2 is a context-centered weakly structured information base system. CM-2 enables accumulation of conceptually promiscuous information from natural information resources. In order for CM-2 information bases to be useful as an information source, we need to develop computational supports for turning conceptually promiscuous information into conceptually well-structured information.

In this paper, we propose several techniques for refining the conceptual structure of CM-2 information bases.

A preliminary version of the proposal is implemented and tested against some information bases from variouse information resources such as travelers' guidebooks, WWW pages.

1 はじめに

個人が所有する情報の量は通信技術の進歩や記憶媒体の大容量化とともに増大する一方であるが、これらの大量の情報を個人が処理する能力には限界があり、計算機による支援が望まれている。

我々はこれまでに連想による概念の関係づけを用いた弱構造の情報ベースシステムとして文脈メディア CM-2 を提案し、手作業で作った CM-2 の情報ベースを、有用で使いやすい構造にするための計算機支援について研究してきた [1]。

本稿では、さらに幅広い情報源から情報ベースを構築するために、ユーザが所有するハイパーテキスト、ドキュメントファイルなどの様々な種類のデータを取り込んで自動的に作った CM-2 の情報ベースを構造化し、ユーザが容易に編集・検索できるようにする手法について述べる。

以下ではまず、文脈メディア CM-2 の表現の枠組について説明する。

次に、自動的にデータを取り込んで作った乱雑な情報ベースと、使いやすい形に構造化された情報ベースを比較し、構造化の目的を明らかにする。

さらに、この乱雑な CM-2 の情報ベースを手作業で構造化する過程を分析して構造化に必要な操作を抽出し、これに基づいて設計した構造化支援パッケージについて説明する。

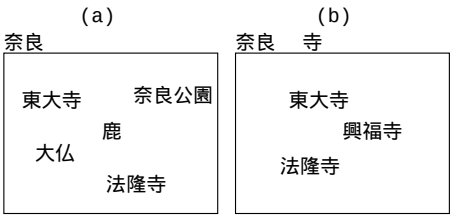

Fig 1: 連想関係の例

最後に、実際にWWW のページからデータを CM-2 の情報ベースに取り込み、構造化を行なった実験結果を示し、本研究のアプローチに関連する問題について議論する。

2　文脈メディア CM-2 の連想表現

CM-2 の情報ベースはユニットと呼ばれる構成要素からなる。ユニットは概念、外部参照データ、連想関係を表すのに用いられる。概念ユニットは情報ベースで用いられる概念を表すユニットで、ユーザの理解のために 1 つのラベルを持つ。連想関係ユニットは、与えられた概念の集合 (以下 **key** ユニットと呼ぶ) と、そこから自然に想起される概念、または外部参照データの集合 (以下 **value** ユニット) を結び付ける。例えば Fig 1は、奈良の観光ガイドから手作業で作った情報ベースの一部である。 (a) の連想表現は “奈良” という概念を “東大寺”、 “法隆寺”、 “奈良公園”、 “鹿”、 “大仏” という概念名が入った箱で図的に表現したものである。外部参照データユニットは実際のデータへのポインタで、画像、テキストなどを扱う。

複数の概念からの連想関係を表現するために、連想関係ユニットは複数の key ユニットを持つことができる。概念 $k_1, \ldots, k_n$ を key ユニットとして持つ連想関係ユニットは、これらの概念から共通に連想される概念の集合を value ユニットとして持つ。例えば Fig1(b) は、 “奈良” と “寺” という 2 つの key ユニットが与えられた時、 “東大寺”、 “興福寺”、 “法隆寺” が連想されることを示している。

3　情報ベースの構造

本研究では、様々な形のテキストメディアからデータを連想表現の形で情報ベースに取り込み、これを徐々に構造化して、ユーザが使いやすい構造を持った情報ベースに近づけていく (Fig 2) 。

Fig 3の左上の連想表現は初期状態の情報ベースの一部であり、その他の連想表現は目標状態の情報ベースの一部である。左上では一つの連想関係の値に電話番号や場所などの様々な種類の概念が入っているという点で乱雑である。また目標状態の情報ベースでは、電話番号や場所が異なる

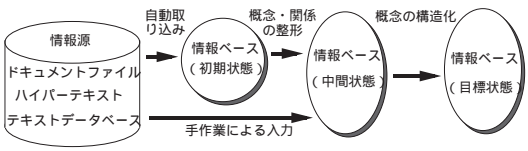

Fig 2: 情報ベースの構造化

連想関係で結ばれている。

- それぞれの連想表現の value ユニットが均質な概念で構成されている。ここで概念が均質とは、 (1) 概念の種類が同じで、 (2) 概念の詳細度が同等であることである。
- 異なる連想関係は異なる種類の関係を表すようになっている。

本節では情報ベースの構造の初期状態と目標状態について分析し構造化の指針を立てる。

3.1　情報ベースの目標状態

ユーザが情報ベースを使う直接の目的はデータや概念の検索である。しかし単に思いつきで何かのデータを検索するわけではなく、何らかの知的作業を達成する手段として検索を行なうことが多い。例えばユーザが情報検索を行なう動機として、

- 論文のサーベイをするために論文情報の空間から類似した論文を集めてきたい
- 旅行のプランを立てるために観光地の見所や移動時間、入場料などを知りたい
- ある分野について勉強したいので理解の助けとなる資料が欲しい

などが考えられる。

知的作業を行なうための検索では、検索対象を明確に表現できないような場合が多い。また検索したデータだけでなく、関連データや類似データなども参照したいことがよくある。

情報ベースをユーザビリティの側面からとらえると、次のような点が情報ベースを設計するうえで重要になる。

- 検索容易性：例えば図書館で所望の本を探す場合に司書に著者と書名を告げてその本が置いてある書架を教えてもらうように、計算機がそのデータベースの構造を理解していて、ユーザが構造を知らなくても、適切な指示を与えればワンショットで検索ができる。
- 理解容易性：しかし、ユーザが検索方法を知らなかったり、探したいものをはっきりと表現できないような場合には前述の方法では所望のデータに到達すること

は難しい。このような場合、さきの図書館の例でいえば、図書館に良く通う人はその図書館での本の分類の仕方を理解していて、ありそうな場所を自分で見つけるかも知れない。

すなわち、ユーザが情報ベースの構造をある程度理解しているならば、ワンショットでは到達しないかもしれないが、その構造の上をたどっていくことで、曖昧にしか表現できないような対象でも容易に検索できる。またこのような検索を行なっているうちにユーザはさらに情報ベースの構造に対する理解を深めることができる。

さらに情報ベースに格納されている情報がある分野の教科書や専門書のようなものであれば、情報ベースの構造の理解は直接その分野の学習、理解につながる。

- 関連データへのアクセス容易性：図書館では見つけ出した本のまわりにも興味深い文献が並んでいることが多い。これは同じカテゴリーに分類された図書をおなじ書架に集めてあるからである。

  情報ベースにおいても、関連のあるデータが同時に表示されたり容易にアクセスできるようにすることで、知的な情報活動を促進する。
- 構造の修正容易性：図書館のように良く整理されているようなところでも、自分では関連していると思っている本が遠くはなれた書架に保管されているようなことがしばしばある。また、自分の本棚を図書館の十進分類法で整理している人は少ないであろう。

  このように、他人が決めた固定された構造がユーザにとって理解しにくかったり、個人の持つ情報の量、種類によって適切な整理の仕方が異なる場合がある。情報ベースの内容、利用するユーザによって検索、理解しやすいようにその構造を修正できる仕組みが必要である。

3.2 CM-2 の情報ベースの初期状態

CM-2 の情報ベースは概念や関係の種類を陽に定義せず、ユーザの自由な概念記述を許している。これによって、ユーザが初めは何を概念にしたらいいのか、概念の関係をどのように表せばいいのか分からくても情報ベースを作ることができるようになっている。ユーザは情報ベースを利用したり編集したりする過程で徐々に望ましい概念の記述方法を習得していく。

このため、ユーザが初めに作った情報ベースや、自然情報源から機械的に取り込んで作られた情報ベースは初めから前述の条件を満たすような使いやすい形をしていない。

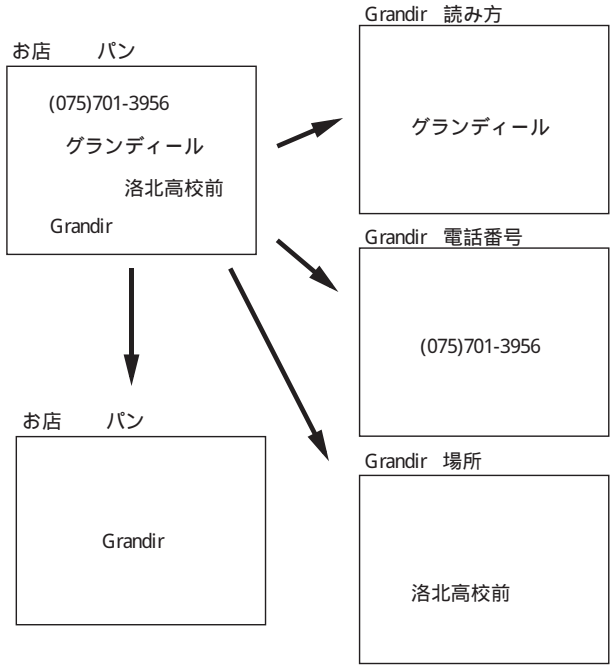

Fig 3: 情報ベースの乱雑さ

例えば Fig 3左上の連想表現は WWW のページから機械的に作った情報ベースの 1 部である。この連想関係は key ユニットとの関係が異なる value ユニットが混在しているという点で乱雑である。他には、あまりにも一つの連想関係ユニットあたりの key ユニット、または value ユニットの数が多いために理解しにくいものや、全く関係がないのに連想関係になっている場合もある。また、概念に関しても、情報源メディアに特有の表現や、用いている自然言語処理技術の限界（例えば方言や未定義語など）、人間が入力するときのタイプミスなどで、 CM-2 の情報ベースを構成する概念としては不適切と思われるものも多く取り込んでしまう。

構造化の作業はこれらの不適切な概念に修正を施すことから始まり、その後で連想関係の修正を行なう。

4 CM-2 の情報ベースの構造化

4.1 手作業による情報ベースの構造化

取り込んだ乱雑な情報ベースを構造化する過程を観察すると、いくつかの共通な操作があることがわかる。

- 概念の修正
  - DELETE-CONCEPT: 不適切な概念の削除
  - REPLACEE-CONCEPT: より一般的な用語での置き換え
  - DIVIDEE-CONCEPT: より一般的な概念への分割 (DIVIDE-ASSOCIATION を伴う)
  - MERGEE-CONCEPT: 複数の概念の集合を 1 つの概念に概念化

- GENERATEE-CONCEPT: 新たな概念を生成

- 連想関係の修正
  - DELETE-ASSOCIATION: 不適切なリンクの削除
  - DIVIDE-ASSOCIATION:DIVIDE-CONCEPT に対応する
  - MERGE-ASSOCIATION: 不必要なリンクを一つにまとめる
  - GENERATE-ASSOCIATION: 新たなに概念にリンクを張る

手作業による構造化の観察により抽出された基本的操作のうち、その操作が行なわれる状況と、操作の具体的なアルゴリズムが記述できるものに関しては、何らかの計算機支援が可能である。特に情報ベース中からある条件を満たしている概念ユニットや連想関係ユニットを捜し出してくるようなタスクに対しては強力な手段となる。

4.2 構造化作業の計算機支援

CM-2 では情報ベースの構造化を支援するためのヒューリスティックの集合である構造化支援パッケージを用意している。

- 同名概念統合ヒューリスティック：同じラベルを持つ概念ユニットを統合することを提案する。操作のレベルでは概念 MERGE の一種と考えられる。
- 同一文脈統合ヒューリスティック：同じ key ユニットから連想される value ユニットを統合した新たな連想関係ユニットを生成することを提案する。拒否された場合は、その違いを説明する新たな key ユニットを付加することを要求する。関係 MERGE によって行なわれる。
- 詳細構造優先ヒューリスティック：ある value ユニットを連想する連想関係ユニットが複数ある時、より詳しい key ユニットから連想されるものを優先することを提案する。関係の DELETE に対応する。
- 説明的連想生成ヒューリスティック：ある概念ユニットのラベルが他の概念ユニットのラベルを含んでいる時、その概念ユニットにラベルが含まれる概念ユニットの集合との間に連想関係を生成することを提案する。さらに、その概念ユニットが key ユニットの要素である場合、ユニットの集合で置き換えることを提案する。操作のレベルでは前者は関係の GENERATE、後者は概念の REPLACE である。
  ユニット x のラベルを $L[x]$ で表すと、

  if $L[x]$ and $L[y]$ are included in $L[z]$

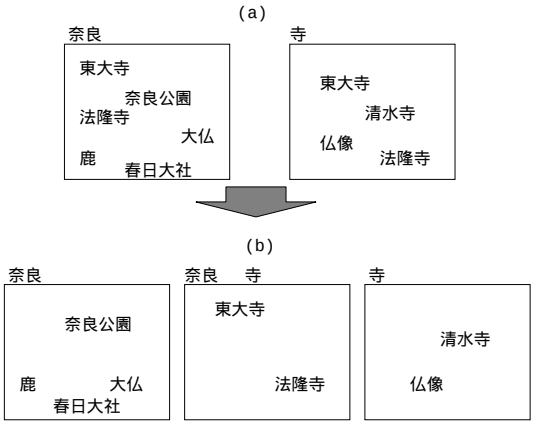

Fig 4: 直交分解

  then suggest to generate $V[\{x\ y\}] = z$

  if $V[z] \neq \{\}$ **then** suggest to replace z by $\{x\ y\}$.

例えば、このヒューリスティックは、“興福寺の五重塔”、“興福寺”、“五重塔” という概念ユニットがあった時に、$V[\{$ “興福寺” “五重塔” $\}] =$ “興福寺の五重塔” 、という連想関係を提案し、さらに $V[$ “興福寺の五重塔” “創建” $] =$ “730 年” という連想関係があった時、これを $V[\{$“興福寺” “五重塔” “創建”$\} =$ “730 年” に置き換えることを提案する。

- 同義語探索ヒューリスティック：情報ベース中での概念の連想関係の構造から類似関係にある概念を探しだし、同義語でないかどうか確認する。

  if $\mathrm{Sim}[a, b] > \theta$, and $\neg\exists x[\ a \in \mathrm{V}[x] \wedge b \in \mathrm{V}[x]\ ]$

  then suggest that a and b are synonyms and assert $b \in \mathrm{V}[\{a$ “synonyms”$\}]$ and $a \in \mathrm{V}[\{b$ “synonyms”$\}]$.

- 直交分解ヒューリスティック：上位クラス形成ヒューリスティックとは反対に、2 つの連想関係の key ユニットのインタセクションを作ることで、与えられた情報ベースの構造を均質な連想関係の集合にしていく。

  if $x \in \mathrm{V}[a_i]\ \ (i = 1, \ldots, n)$ **then** suggest $x \in \mathrm{V}[\{a_1 \cdots a_n\}]$.

例えば、 Fig 4(a) のような連想関係がが与えられたとき、“東大寺” と “法隆寺” という概念は “奈良”、“寺” の両方の key ユニットから連想されているので、直交分解ヒューリスティックは Fig 4(b) に示すように “奈良” と “寺” を key ユニットとする連想関係を新たに作り、“東大寺” と “法隆寺” がその value ユニットとな

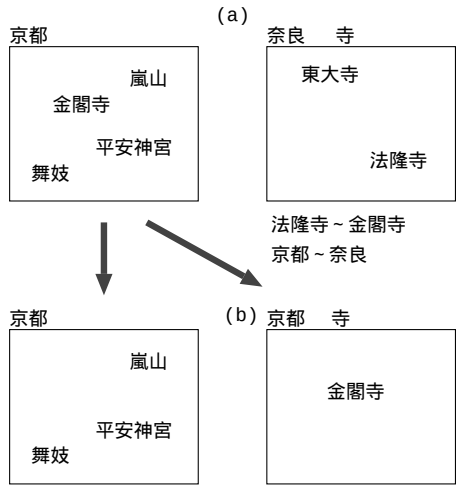

Fig 5: 類推的詳細化

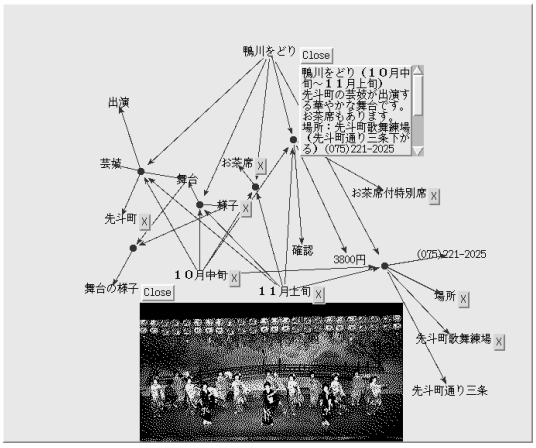

Fig 6: CM2 のユーザインタフェース

る。

- 類推的詳細化ヒューリスティック：類似度を用いてより詳細な連想関係を類推し、さらに情報ベースを構造化する類推的詳細化は類似度に基づいて概念を詳細化するが、類似度の計算には 2 つの概念間のリンクの構造に関する類似度のみを用いている。

$$\textbf{if } x \in \mathrm{V}[p],\ y \in \mathrm{V}[\{qa\}],\ p \neq q, \text{ and } x \notin \mathrm{V}[\{pa\}]$$
$$\textbf{then } penalty \leftarrow \frac{\mathrm{Sim}[x,y]+\mathrm{Sim}[p,q]}{2}$$
$$\text{and}$$
$$suggestion \leftarrow \text{add } x \text{ to } \mathrm{V}[\{pa\}]\ .$$

類推的詳細化ヒューリスティックは、直交分解のように直接その概念を想起する概念が key ユニットになくても (この場合は “金閣寺” という概念が “寺” という概念から連想されていなくても)、概念の類似関係を参照することでより詳細な概念構造を類推することができる。

これらのヒューリスティックの、同名概念統合ヒューリスティック、同一文脈統合ヒューリスティック、詳細構造優先ヒューリスティック、直交分解ヒューリスティック、類推的詳細化ヒューリスティックについてはインプリメントが完了しており、直交分解ヒューリスティック、類推的詳細化ヒューリスティックの有効性については評価実験を行なっている [1]。

4.3　情報ベースの構造化のアルゴリズム

手作業による構造化　4.1の操作を CM2 の編集機能を用いて手作業で行なう。 CM2 の編集機能には (1) 連想関係の削除、生成、つなぎ替え、(2) 概念の削除、生成、(3) キーワード検索、(4) 連想検索機能 [2]、(5) 概念、外部データの参照などの機能が用意されている。具体的には、情報ベースの作業空間で、ユーザが概念や連想関係に不自然さを感じる時や、もっと情報を入れたい時、自分の知識や外部データなどを参照しながら概念を編集する。その編集の結果はその作業空間だけでなく情報ベース全体に反映されるので、異なる作業空間でまた同じ作業をする必要はない。

Table 1: 概念の修正の実験結果 (概念ユニット数:1261、連想関係ユニット数:459)

修正の種類	DELETE	REPLACE	DEVIDE	MERGE	計
修正概念数	166	50	135	23	374

計算機支援を用いた構造化　構造化支援エージェントが作業空間の中に住んでおり、ユーザとエージェント、エージェント同士が協調作業を行なうことで構造化を進める。エージェントはまず CM-2 の情報ベース中で連想関係の乱雑さに不整合を見つけると、それぞれの連想にペナルティを与える。次に乱雑さを解消するヒューリスティックを選定し、ペナルティの大きいものからユーザに構造化するかどうかの確認を行なう。返答が “YES” のとき情報ベースを修正を行なうが今後の構造化のために “NO” の返答も記憶しておく。

5　実験と検討

5.1　概念ユニットの修正に関する実験

京都の観光に関する WWW のページから全自動で取り込んだ弱構造情報ベースに対して、 CM-2 の検索・編集機能を用いて手作業で概念ユニットの修正を行なった。 Table 1は 1261 の概念ユニットのうちいくつの概念ユニットが修正されたかを示している。ちょうど 3 分の 2 の概念は機械的に取り込んだ概念でも概念自体には修正を加えなくても情報ベースとして使えることが分かった。しかし 30% のノイズ

Table 2: 構造化の実験結果 (修正あり)(概念ユニット数:1849、連想関係ユニット数:862)

構造化の種類	適切	不適切	評価不能	総候補数
直交分解	79	121	12	212
類推的詳細化	20	45	0	65
計	99	166	12	277
率	36%	60%	4%	100%

Table 3: 構造化の実験結果 (修正なし)(概念ユニット数:695、連想関係ユニット数:231)

構造化の種類	適切	不適切	評価不能	総候補数
直交分解	229	900	0	1129
類推的詳細化	12	217	0	229
計	241	1117	0	1358
率	18%	82%	0	100%

率は構造化には大きなノイズを与えると予想される。次の 2 つの実験がそれを示している。

5.2 構造化に関する実験 1

人手によって作られた奈良の観光に関する情報ベースに同名概念統合、同一文脈統合、詳細構造優先のヒューリスティックを施してできた約 1800 の概念からなる情報ベースに対して、直交分解と類推的詳細化のヒューリスティックをそれぞれ適用し、評価を行なった。 Table 3はヒューリスティック適用の結果、構造化の候補として提案された概念に対して、ユーザが適切、不適切、評価不能の評価を行なった結果を示している。

直交分解では、 79 の適切な詳細化が起こった。直接の想起関係がない概念についても、類推的詳細化によってさらに構造化された 20 の連想関係が作られた。

5.3 構造化に関する実験 2

WWW から自動収集した関西の観光データについて、事前に概念や構造の修正を行なわず、直交分解と類推的詳細化のみを行なった。 Table 3 のように、事前に修正をしない場合に比べて構造化の精度が大きく落ちている。

6 おわりに

本稿では、自然情報源から取り込んだ乱雑な情報から情報ベースを構築する手段として、文脈メディア CM-2 の情報ベースを、人間と計算機の協調作業によって漸進的に構造化を行なう手法を提案した。そして生の情報ベースと構造化された情報ベースの観察をもとに設計した構造化ヒューリスティックについて説明した。試作バージョンの作成を行ない、 CM-2 の情報ベースに、直交分解と類推的詳細化の 2 つのヒューリスティックを適用する際、事前の概念、連想関係の修正が後の構造化に影響を与えることを示した。

参考文献

[1] 糀谷和人, 武田英明, 西田豊明. 文脈メディアによる情報ベースの構造化. 1995 年度人工知能学会 (第 9 回) 論文集, pp.399–402.

[2] 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 (HIS'95 シンポジウムにて発表予定.)

著者略歴

糀谷 和人
1994 年 大阪大学工学部電子工学科卒業。
同年 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入学、現在に至る。

前田 晴美
1986 年 京都大学文学部哲学科 (心理学専攻) 卒業、同年富士通株式会社入社。
1992-1993 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。
1994 年 MSc. in Computation。
1995 年 富士通株式会社退社、同年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入学、現在に至る。

西田 豊明
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。
1980 年 同大学博士課程退学、同年より京都大学工学部助手。
1988 年 助教授。
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。
1984 年 1 年間 Yale 大学客員研究員。
1995 年 科学技術庁金属材料研究所技術研究所客員研究官。
京都大学工学博士。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。
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) 等。