文脈メディアによる情報ベースの構造化

Refining Information Base with a Contextual Medium

糀谷 和人　武田 英明　西田 豊明

Kazuto Koujitani　Hideaki Takeda　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

The contextual medium CM-1 is a context-centered weakly structured knowledge medium. CM-1 enables accumulation of conceptually promiscuous information from natural information resources. In order for CM-1 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 a couple of techniques for refining the conceptual structure of CM-1 information bases. *Orthogonal decomposition* allows for reorganizing a given information base into a Cartesian product of primitive concepts. *Analogical refinement* enables further elaboration based on the measurement of similarity. A preliminary version of the proposal is implemented and tested against a small information base containing about 1800 units.

1　はじめに

マルチメディア技術などの発達により、様々な種類の情報を様々な形で手に入れることができるようになってきているが、その中でもテキストメディアはもっとも有用な情報源である。

従来のように、初めからデータを意味ネットワークや述語論理のような知識表現で入力して情報ベースを作るアプローチは、知識ベース構築者に多大な労力を要し、非常に限られた範囲を対象にした知識ベースしか作ることができなかった。また、さらに汎用な知識を獲得しようとすると、データが古くなって使い物にならなくなってしまうという問題があった。

世の中に存在する自然言語テキストやハイパーテキストなどの広い範囲のデータを構造化し情報ベースを構築するには、まず弱い構造の枠組に情報を取り込み、人間と計算機の協調作業で情報ベースを少しずつ構造化して、意味ネットワークや述語論理のような強い構造を持った知識表現へ近づけていくというアプローチが有効である。

我々は連想による概念の関係づけを用いた弱構造の枠組として従来から文脈メディア CM-1 を提案してきた [1]。

本稿ではこの文脈メディア CM-1 の情報ベースを、ヒューリスティックを用いてどの程度まで構造化できるかを中心に述べる。

以下ではまず、文脈メディア CM-1 の表現の枠組について説明する。次に、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用いて、概念的に乱雑な情報ベースが均質な情報ベースに構造化される可能性を示す。最後に、それぞれの規則に対する評価実験の結果を示し、本研究のアプローチに関連する問題について議論する。

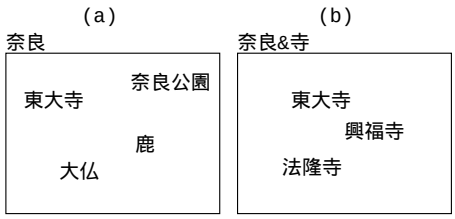

図 1: 文脈の例

2　文脈メディア CM-1

CM-1 の情報ベースはユニットと呼ばれる構成要素からなる。ユニットは文脈と概念を表すのに用いられる。文脈は、与えられた文脈の名前から自然に想起される概念の集合である。例えば図 $1(\mathrm{a})$ の文脈は “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。

複合した概念と文脈の名前を作るために、概念レベルの共通集合オペレータ ‘&’ を用いる。名前 $c_1, \ldots, c_n$ が与えられたとき、 $c_1 \& \ldots \& c_n$ は、 $c_1, \ldots, c_n$ が表す概念または文脈によって共有される概念または文脈を表すのに用いられる。例えば図 $1(\mathrm{b})$ では、奈良にある寺を表すのに “奈良” & “寺” というシンボルを用いている。

CM-1 の情報ベースは、次のような方針によって構築されている。

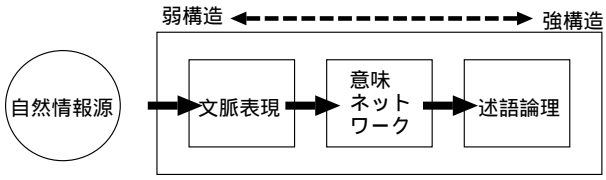

図 2: 情報ベースの構造化

- 概念 - 属性 - 値、クラス - インスタンス、スーパークラス - サブクラス、全体 - 部分などの概念関係を区別しないで、文脈 - 構成要素、概念的共通集合という単純な関係に縮退させる
- 用語や概念に関して標準化をしない
- 人間の主観的な決定を許す
- 概念の矛盾を許す

本研究では、自然言語テキストやインタビューで獲得した知識などの自然情報源を文脈表現の形で情報ベースに取り込み、これを徐々に詳細化して、意味ネットや述語論理のような強い構造を持った情報ベースに近づけていく (図 2) 。

3 CM-1 の情報ベースの構造化

3.1 情報ベースの構造化

自然情報源から取り込んで作られた情報ベースは乱雑な構造をしている。例えば図 $1(\mathrm{a})$ の “奈良” という文脈は、様々な種類の概念が混在しているという点で乱雑である。それに比べて図 $1(\mathrm{b})$ の文脈はより均質な概念で構成されている。CM-1 の情報ベースに文脈の乱雑さに不整合が見つかると、それぞれの文脈にペナルティを与え、乱雑さを解消するヒューリスティックをペナルティの大きいものから適用していく。現在、文脈の形で表現された乱雑な情報ベースを均質なものにするための 2 つの構造化ヒューリスティックを設計し、実験、評価を行なっている。

- 直交分解：文脈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を均質な文脈の集合に分割する
- 類推的詳細化：類似度を用いてより詳細な文脈の構造を類推し、さらに情報ベースを構造化する

4 CM-1 の情報ベースの直交分解

4.1 文脈の直交性

CM-1 では、与えられた文脈の識別子が概念レベルの共通集合オペレータ ‘$\&$’ によって作られたものでないとき、それをプリミティブな文脈と呼ぶ。概念 x に対して、 x がプリミティブな文脈のとき $\{x\}$ を、 $x = y\&z$ のとき $\mathrm{I}(y) \cup \mathrm{I}(z)$ を、 x の文脈インデクスと呼び、 $\mathrm{I}(x)$ と表記する。 $\mathrm{I}(x) = \mathrm{I}(y)$ のとき、概念 x と y は同一であると考える。概念レベルの共通集合オペレータは対称的、結合的であるので、概念 x, y, z に対して、

$$\mathrm{I}(x\&y) = \mathrm{I}(y\&x), \mathrm{I}((x\&y)\&z) = \mathrm{I}(x\&(y\&z))$$

概念レベルの共通集合オペレータは、非常に単純ではあるが、文脈の階層的な索引づけとして強力な手段となる。

$\mathrm{I}(x) \subset \mathrm{I}(y)$ のとき、文脈 $\mathrm{C}[x]$ は文脈 $\mathrm{C}[y]$ を包摂するという。例えば、

$$\mathrm{I}[\{\text{“奈良”}\}] = \{\text{“奈良”}\}$$
$$\subset \{\text{“奈良”}, \text{“寺”}\} = \mathrm{I}[\text{“奈良”}\&\text{“寺”}].$$

であるので、 “奈良” は “奈良”$\&$“寺” に包摂されることになる。

概念 x が与えられたとき、ユーザは概念 x が自分にとって意味があるかどうかによって重要か重要でない (すなわち無意味) かを決めることができる。ユーザが明示的に $\mathrm{C}[x]$ に要素を与えたときは x は重要であることになる。ユーザは文脈 x の重要度の決定を保留することもできる。この場合、 x は重要度未知であるという。

文脈 $x_1\&x_2$ が重要であるとき、文脈 x_1 と x_2 は直交するという。例えば図 $1(\mathrm{b})$ のような文脈が与えられたとき、文脈 “日本” と文脈 “都市” は直交することになる。

4.2 直交分解

直交分解は、文脈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を均質な文脈の集合に分割する例えば、図 $3(\mathrm{a})$ のような文脈が与えられたとき、 “東大寺” と “法隆寺” という概念は “奈良”、 “寺” の両方の文脈に含まれているので、直交分解ヒューリスティックは図 $3(\mathrm{b})$ に示すように “奈良 $\&$ 寺” という文脈をつくり、 “東大寺” と “法隆寺” がその中に含まれることになる。具体的には、情報ベース中から 2 つのことなる文脈 $\mathrm{C}[x]$、 $\mathrm{C}[y]$ に共通に含まれる概念 z を発見し、 $\mathrm{C}[x\&y]$ の中にその概念がなければ、この問題を解決するように提案する。

if contexts $x, y \in B$ and context $x\&y$ is not insignificant
then
penalty

$$\leftarrow \frac{\max[|(\mathrm{C}[x] \cap \mathrm{C}[y]) - \mathrm{C}[x\&y]|, |\mathrm{C}[x\&y] - (\mathrm{C}[x] \cap \mathrm{C}[y])|]}{\max[|\mathrm{C}[x] \cap \mathrm{C}[y]|, |\mathrm{C}[x\&y]|]}$$

and *suggestion*

$$\leftarrow \text{resolve the difference between } \mathrm{C}[x] \cap \mathrm{C}[y] \text{ and } \mathrm{C}[x\&y]$$

if contexts $x, y \in B$, $\mathrm{I}[x] \subset \mathrm{I}[y]$, and $\exists z[\, z \in \mathrm{C}[x] \wedge z \in \mathrm{I}[y]\,]$
then

$$penalty \leftarrow \infty$$

and

$$suggestion \leftarrow \text{remove } z \text{ from } x$$

一つめの規則は具体的には、 z を $\mathrm{C}[x\&y]$ に追加することを提案する。

if $z \notin \mathrm{C}[x\&y]$ and $z \in (\mathrm{C}[x] \cap \mathrm{C}[y])$
then *suggestion* $\leftarrow$ add z to $\mathrm{C}[x\&y]$

5 情報ベースの類推的詳細化

5.1 概念の類似度

類推的詳細化は類似度に基づいて概念を詳細化する。 CM-1 の情報ベースは概念 (すなわち文脈) を表すノードと、文脈 - 構成要素、概念的共通集合を表す 2 種類のリンクからなっ

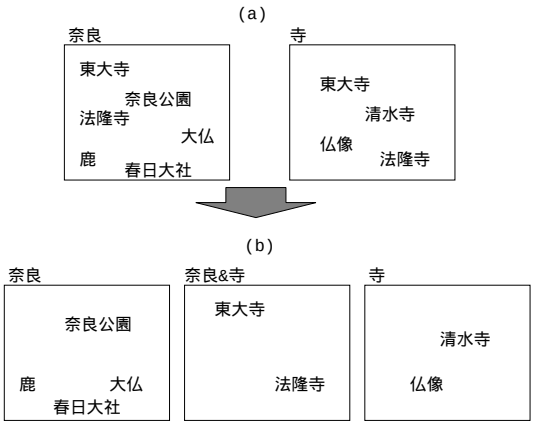

図 3: 直交分解

ている。現在のところ、 2 つの概念の関係構造に関する類似度のみを用いる [2]。

直交しない 2 つの文脈 x、 y が与えられたとき、類似度 $\mathrm{Sim}[x,y]$ を 3 つの視点から計算した類似度 $\mathrm{Sim}^{(\mathrm{a})}[x,y]$ 、 $\mathrm{Sim}^{(\mathrm{b})}[x,y]$ 、 $\mathrm{Sim}^{(\mathrm{c})}[x,y]$ の平均と定義する。

$$\mathrm{Sim}[x,y] = \frac{\mathrm{Sim}^{(\mathrm{a})}[x,y] + \mathrm{Sim}^{(\mathrm{b})}[x,y] + \mathrm{Sim}^{(\mathrm{c})}[x,y]}{3} \in [0,1].$$

$\mathrm{Sim}^{(\mathrm{a})}[x,y]$ は、 $\mathrm{C}[x]$ と $\mathrm{C}[y]$ の中にある概念を比較して x と y の類似性を測る。定義は以下のようになる。

$$\mathrm{Sim}^{(\mathrm{a})}[x,y] = \frac{|\{z \,|\, z \in \mathrm{C}[x] \wedge z \in \mathrm{C}[y]\}|}{|\mathrm{C}[x] \cup \mathrm{C}[y]|}.$$

$\mathrm{Sim}^{(\mathrm{b})}[x,y]$ は x と y の両方を含んでいる文脈の割合である。すなわち、

$$\mathrm{Sim}^{(\mathrm{b})}[x,y] = \frac{|\{z \,|\, x \in \mathrm{C}[z] \wedge y \in \mathrm{C}[z]\}|}{|\{z \,|\, x \in \mathrm{C}[z]\} \cup \{z \,|\, y \in \mathrm{C}[z]\}|}.$$

である。

$\mathrm{Sim}^{(\mathrm{c})}[x,y]$ は x と y の両方に直交している文脈の割合を示している。つまり、

$$\mathrm{Sim}^{(\mathrm{c})}[x,y] = \frac{|\{z \,|\, \langle z \text{ is orthogonal to } x\rangle \wedge \langle z \text{ is orthogonal to } y\rangle\}|}{|\{z \,|\, \langle z \text{ is orthogonal to } x\rangle\} \cup \{z \,|\, \langle z \text{ is orthogonal to } y\rangle\}|}.$$

5.2 類推的詳細化

類推的詳細化ヒューリスティックは、類似度を用いてより詳細な文脈の構造を類推し、さらに情報ベースの概念を分類することを提案する。

例えば図 4(a) の左に示したような乱雑な文脈が与えられたとする。情報ベースの中に右上に示したような構造化された文脈があって、文脈 “奈良” と “京都”、概念 “金閣寺” と “法隆寺” がそれぞれ類似していることが情報ベースから計算されると、これらを参照して図 4(b) の右に示すようなより構造化された新しい文脈を類推して提案する。具体的には、次のような診断ルールによって概念を分類することを提案する。

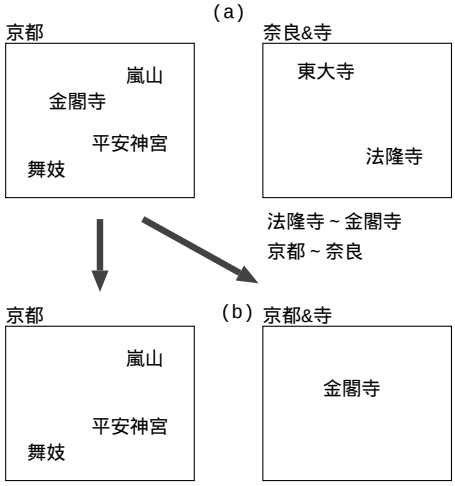

図 4: 類推的詳細化

if $x \in \mathrm{C}[p]$, $y \in \mathrm{C}[q\&a]$, $p \neq q$, and $x \notin \mathrm{C}[p\&a]$

then $penalty \leftarrow \frac{\mathrm{Sim}[x,y]+\mathrm{Sim}[p,q]}{2}$
and
$suggestion \leftarrow$ add x to $\mathrm{C}[p\&a]$.

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文脈がなくても (この場合は “金閣寺” という概念が “寺” という文脈に含まれていなくても) 、概念の類似関係を参照することでより詳細な概念の構造を類推することができる。

6 実験と検討

6.1 実験結果

人手によって無作為に作られた約 1800 の概念からなる奈良の観光に関する情報ベースに対して、これまでに述べた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。表 1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 4.1で説明した重要、無意味、重要度未知の評価を行なった結果を示している。

直交分解では、図 5のような 79 の文脈に関してユーザにとって意味のある詳細化が起こった。図 6のように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 の文脈が作られた。

表 1: 実験結果 (概念数:1849、文脈数:862)

構造化の種類	重要	無意味	重要度未知	総候補数
直交分解	79	121	12	212
類推的詳細化	20	45	0	65

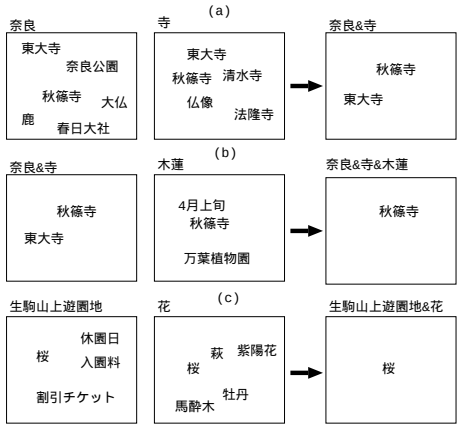

図 5: 直交分解ヒューリスティックの適用結果例

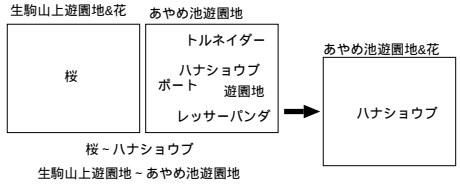

図 6: 類推的詳細化ヒューリスティックの適用結果例

6.2 検討と課題

漸進的な詳細化 図 5の (b) では、 (a) の直交分解の結果にさらに詳細な情報が加わって直交分解が施され、結果としてより均質な文脈が作られている。直交分解が繰り返されることで乱雑な概念がより詳細で均質な概念に構造化されていくことを示している。また図 6は図 5(c) で “生駒山上遊園地” の文脈が直交分解の結果詳細化され、 “あやめ池遊園地” との間に詳細度の差ができたため、類推的詳細化がおこり、 “あやめ池遊園地” の方へも詳細化が伝搬している。

このように複数のヒューリスティックを適用していくことで、構造化が縦横両方向に進み、情報ベース全体が構造化されていく。さらに直交分解、類推的詳細化以外のヒューリスティックを開発することで、より構造化が進むことが期待される。

類似度算出の問題点 情報ベースの概念が少ないとき、例えば “駐車場”、 “本尊” という概念が “新薬師寺 & 駐車場”、 “大安寺 & 駐車場”、 “正歴寺 & 駐車場”、 “新薬師寺 & 本尊”、 “大安寺 & 本尊”、 “正歴寺 & 本尊” という文脈でしか使われていないとすると、先に示した類似度の定義より Sim[“駐車場”,“本尊”] $= 1/3$ と非常に高い類似度が計算されるため、不適切な類推的詳細化を起こす可能性がある。

生成した文脈の問題点 現在のところ文脈の直交性を評価するアルゴリズムを実装していないため、例えば図 7のように “休園日 & あやめ池遊園地” と “正暦寺の精進料理 & 定休日” がどちらもたまたま ”水曜日 ”であったために無意味な

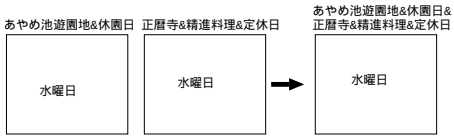

図 7: 不適切な直交分解の例

文脈を生成してしまう。

これに対しては、新たに文脈を生成する際には文脈の直交性を評価するアルゴリズムを組み込む必要がある。これを解決するための一つの方法として概念のクラス情報を用いることが考えられる。

十分構造化されて均質になった文脈はインデックスとしてその文脈を表すクラス概念を含んでいると考えられる。すなわち、外部から明示的にクラス情報を含んだ形でデータを獲得しないでも、文脈の重要度を評価することによって概念の isa 階層を作ることができる。例えば図 3(b) の真中の文脈では “寺” というインデックスがこの文脈のクラス概念を表している。そこで重要である文脈の要素に対して、ユーザにそのクラスを文脈のインデックスから選択させるヒューリスティックを導入することを考える。

if C[x] is significant
then ask which of I[x] describes the class of the context

また、この isa 階層を用いることで、 “寺 & 尼寺”、 “本数 & 数” のようなクラス - サブクラスの関係にある概念や、 “法華寺 & 中宮寺 & 円照寺” のような同じクラスの概念をインデックスにもつ、直交性の少ない文脈の生成を防ぐことができる。

これについては将来さらに研究を行なう余地がある。

7 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-1 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 CM-1 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ことで、漸進的な詳細化が可能であることを示した。試作バージョンの作成を行ない、 1800 の構成要素からなる情報ベースに対してテスト行ない、 2 つのヒューリスティックの有効性を示した。

参考文献

[1] Nishida, T. and Takeda, H.: Towards the Knowledgeable Community, In Fuchi, Kazuhiro and Yokoi, Toshio (eds.)*Knowledge Building and Knowledge Sharing* , pp. 155–164, IOS Press, 1994

[2] G. Bisson. Why and How to Define a Similarity Measure for Object-Based Representation Systems. In N.J.I. Mars, Ed. *Towards Very Large Knowlege Bases: Knowledge Building & Knowledge Sharing 1995*, pp.236–246. IOS Press, 1995.