ジャンル体系に基づく概念辞書構築手法の提案

佐藤 吉秀† 安部 伸治† 大久保 雅且†
† NTT サイバーソリューション研究所

1 はじめに

著者らは，関連するコンテンツの連鎖的な発見を支援するガイドシステム “AssociaGuide[1]” の研究を進めている．ユーザは，2 次元のマップ上に配置されたコンテンツを俯瞰し，視点を連続的に変化させながら散策してコンテンツへ到達する．コンテンツの配置は，内容の近いコンテンツが近くに配置されるよう相互の類似度に基づいて行うが，それと同時に，コンテンツを分類するジャンルをクラスタで表した配置を行うため，全コンテンツの俯瞰から特定のジャンルへのズームイン，コンテンツへの到達，関連するコンテンツの発見までの流れを，思考を中断させることなく行うことができる．

類似度の判定には概念辞書と呼ぶ知識ベースを使用するが，ジャンルを意識したコンテンツ配置を行うには，概念辞書自体にジャンル構造を持たせることが必要である．本稿では，ジャンル構造を持つ概念辞書を自動構築するための 1 手法を提案する．

2 ジャンル体系に基づく概念辞書

コンテンツの類似度判定に用いる概念辞書は，表 1 のように各語句を複数の概念に対する属性値列 (ベクトル) で表現した行列である．「馬」や「飛行機」といった語句の概念が，基底語と呼ぶ「生物」「器官」などの概念との関連度を数値化した属性値列，すなわちベクトルで表されている．このため，語句の概念が類似しているほどベクトル間の距離が小さい．概念辞書は,「馬」や「飛行機」といった語句が分布した高次元空間とみなすことができ，基底語は高次元空間の軸に対応する．

表 1: 概念辞書

語句	基底語				
	生物	器官	機械	乗物	···
馬	0.4	0.2	0.1	0.1	···
飛行機	0.0	0.0	0.5	0.3	···
⋮	⋮	⋮	⋮	⋮	

概念辞書が形成する高次元空間における語句の分布は，基底語の選び方の影響を大きく受ける．図 1 は，基底語の一般性の強弱による空間中の語句分布の変化を表した模式図である．各語句はあらかじめ定めた複数のジャンルのうち，いずれか 1 ジャンルに属するものとする．(A) のように，ある特定のジャンルに強く関連する語句を基底語に選んだ場合，異なるジャンルに属する語句のベクトル同士が直交しやすくなり，各ジャンルが独立した空間となる．逆に，(C) のように汎用性の高い語句を基底語に選ぶと，異なるジャンルに属する語句のベクトルが著しく混在してしまう．ジャンル体系の存在を前提とする類似度判定に適した概念辞書とは，(A) と (C) の中間的な，(B) のような空間を形成する概念辞書である．

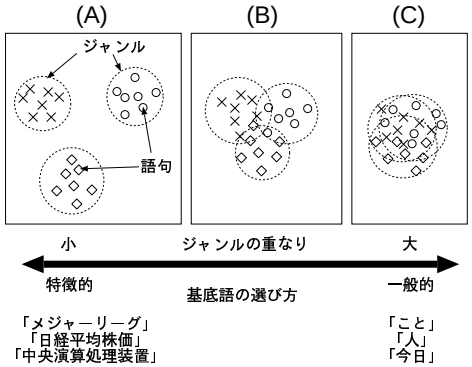

図 1: 基底語の選び方と高次元空間中の語句分布

3 空間情報量による概念辞書の最適化

概念辞書が図 1 の (A) や (C) のような空間を形成している場合，空間内の語句の配置には無駄があり，空間全体が与える情報量の損失が大きい．これらの概念辞書は，いずれもジャンル構造が適切に表現されず，ジャンル体系の存在を意識したコンテンツ配置には適さない．

自動的に (B) のような空間を形成する概念辞書を構築するため，空間全体から得ることのできる情報量を表す空間情報量を定義し，空間情報量に基づいて概念辞書を最適化する方式を提案する．

まず，図 2 において，異なる 2 ジャンル G_x, G_y に属する語句が空間上で混在する割合を表すジャンル重複率を次式で定義する．

$$P(x,y) = \frac{\sigma_{xy}}{\sigma_x + \sigma_y - \sigma_{xy}} \quad (1)$$

The method of constructing concept dictionary based on directory tree
†Yoshihide Sato, Shinji Abe and Masaaki Okubo
NTT Cyber Solutions Laboratories

σ_x, σ_y は，各ジャンルに属する語句の重心ベクトルからの偏角の標準偏差である．重心ベクトルからの偏角が標準偏差以下であるような範囲をジャンルの領域と呼べば，これらは各ジャンルの領域の半径とみなせる．σ_{xy} は，2 ジャンルの重複領域 G_{xy} に含まれる語句について求まる標準偏差である．

続いて，空間情報量 I_{space} を，異なる 2 ジャンル G_x, G_y 間のジャンル重複率から導かれる情報量 $\log\{1/P(x,y)\}$ [1] の期待値として，式 (2) で定義する．

$$\begin{aligned} I_{\mathrm{space}} &= \sum_{x,y} P(x,y)\log\{1/P(x,y)\} \\ &= -\sum_{x,y} P(x,y)\log P(x,y) \end{aligned} \quad (2)$$

4 実験

高次元空間中の語句の分布に大きな影響を与える基底語の選び方ついて調べるため，多数の文書中から抽出した語句集合の部分集合として複数の基底語セットを作成し，構築される概念辞書の空間情報量の違いを調べた．

現代用語の基礎知識 [3] の解説記事のうち,「経済」「政治」「国際情勢」「各国事情」「情報・メディア」の 5 ジャンルをさらに細分化した合計 47 ジャンル分の解説記事から抽出した 21205 種類の名詞 (句) を利用した．今回は，基底語の一般性の強弱による違いをみるため，汎用性の高い順に並べた名詞 (句) から連続する 500 語を 1 個の基底語セットとし，これを複数セット作成した．

全 47 ジャンル中 1 ジャンル内の解説記事にのみ出現した語句を対象とし，これらの語句についてベクトルを算出した．ベクトルの要素 (各基底語に対する属性値) は式 (3) にしたがって決定した．$f(w_k)$ は語句 w_k を含む解説記事数，$d(w_m, w_n)$ は語句 w_m, w_n を同時に含む解説記事数であり，$MI(w_m, w_n)$ は，w_m と w_n が同一解説記事中に出現する頻度が高いほど大きな値をとる．

$$\begin{aligned} &MI(w_m, w_n) \\ &= \log \frac{d(w_m, w_n)/\sum_{i,j} d(w_i, w_j)}{\{f(w_m)/\sum_k f(w_k)\}\{f(w_n)/\sum_k f(w_k)\}} \end{aligned} \quad (3)$$

図 3 に，複数の基底語セットついて構築した概念辞書の空間情報量の変化を示す．横軸は基底語の汎用性の強弱を表す．この結果，基底語が特徴的な場合 (A) と一般的な場合 (C) の間で空間情報量を最大にする基底語セットが見つかった．さらに，空間情報量によって最適化した概念辞書 (B) について，高次元空間内の語句を代表的な次元圧縮手法である MDS(多次元尺度構成法) で 2 次元上にプロットした (図 4)，その結果，同一ジャンルに属する語句がクラスタを形成し，かつ異なるジャンルの語句が多少混在しており，主観的ながら図 1(B) のような空間が形成されていることを確認できた．

[1] 一般に，生起確率が $P(a_n)$ で与えられる事象の情報量は $\log(1/P(a_n))$ で与えられる．[2]

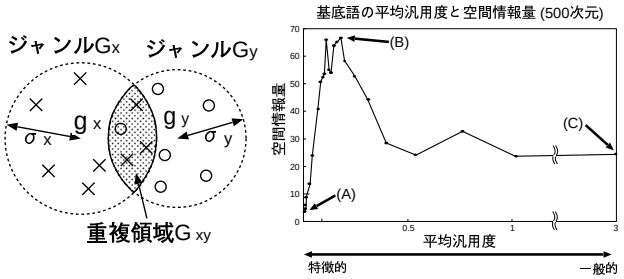

図 3: 平均汎用度と空間情報量

図 2: ジャンル重複領域

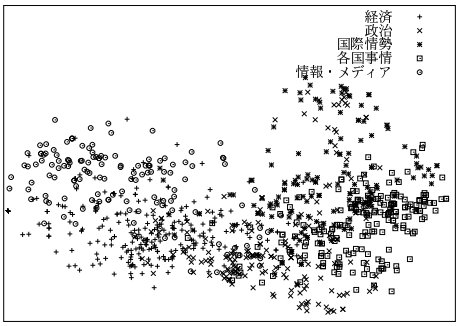

図 4: MDS 適用結果 (空間情報量で最適化した概念辞書)

5 おわりに

ジャンル体系の存在を前提とする概念辞書構築における提案評価手法により，辞書構築の自動化の見通しを得ることができた．空間情報量は，基底語の選び方以外のパラメータに対する評価にも有効である．今後は，提案手法による最適化が実際のコンテンツの配置に与える影響を調べる．

参考文献

[1] 宮原伸二, 藤田悦郎, 安部伸治, 林泰仁: “散策型映像ポータルシステム AssociaGuide の提案”, 2002 年信学会総合大会, D-8-7, (2002)

[2] 瀧保夫, 宮川洋: “岩波講座 基礎工学 19 情報論 I”, 岩波書店, pp.19-20

[3] “システムソフト電子辞書シリーズ 現代用語の基礎知識 2003”, ロゴヴィスタ (株)