サーチングのための言語情報に基づくWebディレクトリのマップ変換技術

藤田 悦郎　安部 伸治　林 泰仁
{fujita.etsuro, abe.sinji, hayashi.yasuhito}@lab.ntt.co.jp
日本電信電話株式会社 NTTサイバーソリューション研究所

1. はじめに

我々は，より付加価値の高いコンテンツナビゲーションサービスの実現に向けた取組みの一環として，コンテンツに付与された意味内容に関するメタデータを活用して，大量のコンテンツを2次元上に分類・マッピングするシステム「AssociaGuide」の研究開発を進めている[1][2][6][7]．本システムでは，インタフェースにマップ表現を採用することにで，ユーザが大量のコンテンツに対し電子地図を操作する感覚で全体像を鳥瞰しながら興味あるコンテンツに連続的にたどりつけることができる．

本システムにおけるコンテンツの分類・マッピング過程では，コンテンツに付随するジャンルなど分類情報と，概要説明文書など言語情報とを統合的に処理することによって，コンテンツ間の意味内容的な関連性と同時に，与えられた分類情報の情報構造を反映した2次元マップを生成する．分類情報の情報構造は，2次元マップにおいて巨視的な構造として陽に表現される．これによってユーザは，2次元マップの意味内容的な構造を概観・把握できるようになるメリットがある．また，連続的なフォーカスのための指標などとして利用することができる．

さらに，コンテンツの分類・マッピング処理では，電子地図のメタファーを考慮する立場から，新規コンテンツの登録については追加型を前提としている．すなわち，新規コンテンツを新たに登録する場合には，それまでに登録されているコンテンツの2次元座標は全く変えずに新規コンテンツのみを2次元マップ上に追加的に配置する．これによってユーザが，本システムを使い続けるなかで，既に閲覧したコンテンツの配置場所を記憶したり，配置場所による意味内容の微妙な差異を視覚的に記憶したりすることができるメリットがある．いわゆる土地鑑を活かした2次元マップの探索・散策ができるようになるのである．これは，インターネットなど，コンテンツが随時追加されるサービスでは特に有用であろう．

以下本稿では，「AssociaGuide」で実現しているメタデータを前提としたコンテンツの分類・マッピング手法を述べ，本手法をWebディレクトリに適用した試みについて報告する．

2. メタデータを前提としたコンテンツ空間の可視化

提案手法では，与えられた分類情報の情報構造，すなわちコンテンツの分類大系を反映した2次元マップ（以下、基準マップと呼ぶ）をテンプレート的にまず生成し，これにコンテンツを一つずつ登録していく．提案手法は，分類大系を参照し，分類大系の最下層ジャンルに対応付けられた一つないし複数の概念ベクトルを用いて基準マップを生成する過程と，入力コンテンツが与えられた場合に，それに付随する言語情報から概念ベクトルを生成し，分類情報を考慮しながら，入力コンテンツを基準マップに追加的に登録する過程からなる．図1に提案システムの概念図を示す．

A 2D Mapping of Web Directories based on Textual Information
Etsuro FUJITA, Shinji ABE and Yasuhito HAYASHI
NTT Cyber Solutions Laboratories

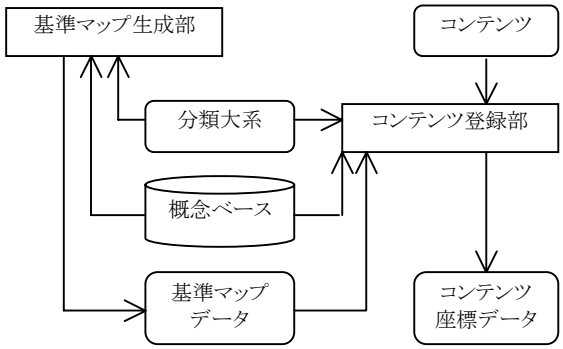

図1：提案システムの概念図

2．1．基準マップの生成

ここでは，分類大系の最下層ジャンルに対応付けられた概念ベクトルを用いて，与えられた分類情報の情報構造を反映した基準マップを生成する．基準マップは，すべての最下層ジャンルに対応付けられた概念ベクトルから，それら概念ベクトルどうしの概念空間での距離関係を考慮しつつ，それら概念ベクトルが属するジャンルの深さ方向に関する一致度合いを同時に勘案して，深さ方向に一致すればするほど，概念ベクトルどうしが2次元上で互いに近い位置に配置されるようある制約を設けて配置する．これによって，1階層目が同一ジャンルの概念ベクトルどうしは，2次元上で互いに近い位置に配置されることになる．さらに，2階層目も同じジャンルに分類される概念ベクトルどうしは，そのなかでもさらに近接して配置されることになる．そして，このような分類情報の階層構造に対応した概念ベクトルのクラスタ化が最下層ジャンルまで続く2次元マップが生成されることになる．一方，基準マップでは，概念ベクトルどうしの概念空間での距離関係を考慮するため，各階層におけるジャンルのクラスタの周囲には，内容的に近いジャンルのそれが配置されることになる．したがって，分類情報の階層構造と，ジャンルどうしの意味内容的な関連性とを同時に反映した2次元マップが生成されることになる．なお，ここで述べた概念ベクトルは，日本語語彙体系における約3000の意味カテゴリに対する関連度を成分とする多次元ベクトルを意味している[3][5]．

以下，基準マップの生成アルゴリズムについて詳説するが，ここでは簡単のため，分類情報の階層構造が(すべての枝で)深さ2の場合を例に説明する．提案アルゴリズムは，深さが任意の場合にも容易に拡張可能である．また，枝によって深さが異なる場合にも拡張可能である．

1階層目のジャンルを$G_p(p=1,\cdots,N_{ROOT})$とし，G_pに含まれる2階層目のジャンルを$G_{pq}(q=1,\cdots,N_p)$とする．ここで，N_{ROOT}は1階層目のジャンル数を，N_pはG_pに含まれる2階層

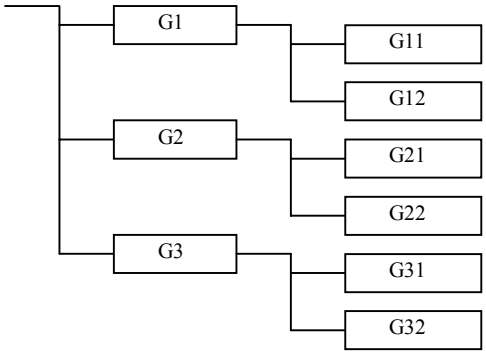

図2：分類大系の例

目のジャンル数を表す．図2に分類大系の例を示す．また，図3に，図2の場合に対応する基準マップの概念図を示す．

2階層目のジャンル G_{pq} に対応付けられた概念ベクトルの集合を S_{pq} とする．また，1階層目の各ジャンル G_p について S_p を次のように定める．また，S を次のように定める．

$$S_p = S_{p1} \cup \cdots \cup S_{pN_p} \tag{1}$$

$$S = S_1 \cup \cdots \cup S_{N_{ROOT}} \tag{2}$$

S は2階層目のすべてのジャンル G_{pq} に対応付けられた概念ベクトル全体の集合である．

基準マップの生成では，S に含まれる概念ベクトルを，それら概念ベクトルどうしの概念空間での距離関係が保存されるように多次元尺度法[8]を用いて2次元上に配置するが，その際，前述した分類情報の情報構造を反映するために概念ベクトルの2次元配置に関してある制約を設ける．すなわち，次の目的関数 E を，以下に述べる制約つきで最小化する問題として定式化する．

$$E = \frac{1}{\sum_{i<j} d_{ij}^*} \sum_{i<j} \frac{(d_{ij}^* - d_{ij})^2}{d_{ij}^*} \tag{3}$$

ここで，d_{ij}^* は S に含まれる概念ベクトル v_i および v_j の概念空間でのユークリッド距離を表す．また，d_{ij} は，v_i および v_j に対応する2次元座標 (x_i, y_i) および (x_j, y_j) のユークリッド距離を表す．

以下，制約条件について説明する．まず，1階層目の分類情報の情報構造を基準マップに反映するための制約条件として，$S_p (p = 1, \cdots, N_{ROOT})$ のすべての異なる組合せ S_p および $S_{p'}$ について次の不等式制約を導入する．

$$g_{S_p, S_{p'}}(x_1, y_1, \cdots, x_n, y_n) = d_{S_p, S_{p'}} - \mu_{S_p, S_{p'}}(\sigma_{S_p} + \sigma_{S_{p'}}) \geq 0 \tag{4}$$

ここで，n は S に含まれる概念ベクトル v_i の総数 $\#(S)$ を表す．また，d_{S_p, S_q} は，S_p ($S_{p'}$)に属する概念ベクトルたちに対応する2次元座標の重心を $(\bar{x}_{S_p}, \bar{y}_{S_p})$ (あるいは $(\bar{x}_{S_{p'}}, \bar{y}_{S_{p'}})$)とするとき，これら重心座標間のユークリッド距離を表す．すなわち，$d_{S_p, S_{p'}}$ は

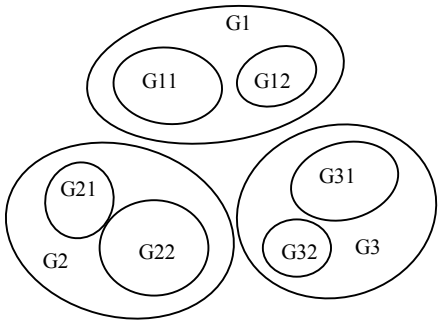

図3：基準マップの概念図

$$\bar{x}_{S_p} = \frac{1}{\#(S_p)} \sum_{v_i \in S_p} x_i \,, \quad \bar{y}_{S_p} = \frac{1}{\#(S_p)} \sum_{v_i \in S_p} y_i \tag{5}$$

とするとき，

$$d_{S_p, S_{p'}} = \sqrt{(\bar{x}_{S_p} - \bar{x}_{S_{p'}})^2 + (\bar{y}_{S_p} - \bar{y}_{S_{p'}})^2} \tag{6}$$

また，σ_{S_p} ($\sigma_{S_{p'}}$)は S_p ($S_{p'}$)に属する概念ベクトル v_i たちに対応する2次元座標 (x_i, y_i) の，重心座標 $(\bar{x}_{S_p}, \bar{y}_{S_p})$ (あるいは $(\bar{x}_{S_{p'}}, \bar{y}_{S_{p'}})$)を中心とする2乗平均の平方根を表す．すなわち，$\sigma_{S_p}$ は

$$\sigma_{S_p} = \sqrt{\frac{1}{\#(S_p)} \sum_{v_i \in S_p} \{(x_i - \bar{x}_{S_p})^2 + (y_i - \bar{y}_{S_p})^2\}} \tag{7}$$

また，式(4)で $\mu_{S_p, S_{p'}}$ は1よりも大きな実数を表す．

式(4)は，S_p および $S_{p'}$ の概念ベクトルたちが2次元上で分離して配置されるようにするための制約である．

次に，2階層目の分類情報の情報構造を基準マップに反映するための制約条件として，p を固定し $S_{pq} (q = 1, \cdots, N_p)$ のすべての異なる組合せ S_{pq} および $S_{pq'}$ について，次の不等式制約を導入する．

$$g_{S_{pq}, S_{pq'}}(x_1, y_1, \cdots, x_n, y_n) = d_{S_{pq}, S_{pq'}} - \mu_{S_{pq}, S_{pq'}}(\sigma_{S_{pq}} + \sigma_{S_{pq'}}) \geq 0 \tag{8}$$

ここで，$d_{S_{pq}, S_{pq'}}$ ，$\mu_{S_{pq}, S_{pq'}}$ および $\sigma_{S_{pq}}$ ，$\sigma_{S_{pq'}}$ は上記と同様にして定義される．この制約式がすべての $p = 1, \cdots, N_{ROOT}$ にわたって導入される．

式(8)の意味も式(4)と同様であり，S_{pq} および $S_{pq'}$ の概念ベクトルたちが2次元上で分離して配置されるようにするためのものである．

式(4)および(8)の不等式制約を同時に満たす2次元座標 (x_i, y_i) の組であって，式(3)を最小化する，すなわち概念ベクトルどうしの概念空間での距離関係を最大限保存するような組を求めることにより，前述した性質を備えた基準マップを生成する．なお，上記の制約つき最小化問題は，逐次2次計画法を用いて解くことができる[4]．

2．2．コンテンツの登録

ここでは，入力コンテンツに付随する分類情報と言語情報から，入力コンテンツを基準マップに追加的に登録する．以下で

は、入力コンテンツに付与された(最下層の)ジャンルが G_{pq} としてアルゴリズムを説明する.

(手順1)コンテンツに付与された概要説明文書あるいはキーワードなど言語情報を概念ベースにかけ, コンテンツの意味内容を特徴付ける概念ベクトルを生成する.

(手順2)ジャンル G_{pq} に対応付けられた概念ベクトル集合 S_{pq} のなかで, 入力コンテンツの概念ベクトルと概念空間でのユークリッド距離が最も近いもの v_k を求める. v_k に対応する2次元座標 (x_k, y_k) を, 入力コンテンツの2次元初期座標 (x, y) とする.

(手順3) (x_k, y_k) を中心とする十分大きな半径 r の円領域 $NB_k(r)$ をとり, $NB_k(r)$ に含まれるすべての (x_i, y_i) を用いて, 次式により入力コンテンツの2次元座標 (x, y) を移動修正する.

$$x' = x + \alpha(r)\sum_i h(d_i^*)[x_i - x] \tag{9}$$

$$y' = y + \alpha(r)\sum_i h(d_i^*)[y_i - y] \tag{10}$$

ここで, $\alpha(r)$ は r に関する単調減少関数であって, $r \to 0$ のとき, $\alpha(r)$ は 0 に十分近い値に減少する. また, d_i^* は入力コンテンツの概念ベクトルと, 概念ベクトル集合 S の要素 v_i とのユークリッド距離を表し, $h(u)$ は u に関する単調減少関数を表す.

この処理によって, 入力コンテンツの2次元座標 (x, y) は, 「概念空間でのユークリッド距離がより近い」2次元座標 (x_i, y_i) の方向に移動させられることになる.

(手順4)手順3の処理結果の (x', y') を (x, y) として, r を徐々に小さくして $NB_k(r)$ を小さくしながら, 手順3の処理を繰り返す. $NB_k(r)$ が (x_k, y_k) 以外の2次元座標を含まなくなったら, この反復処理を終了する.

以上の手続きによって, 入力コンテンツは, 基準マップ上の (x_k, y_k) の付近であってかつ周辺の (x_i, y_i) との「概念空間での距離関係を反映した位置」に配置されることになる. またこの配置処理によって, 意味内容的に近い(すなわち, 概念ベクトル間のユークリッド距離が小さい)コンテンツどうしは, 基準マップ上で互いに近い位置に配置されることになる. なお, $\alpha(r)$ を単調減少させるのは, $r \to 0$ のとき, 式(9)および(10)による反復処理で (x, y) の振動を押さえて収束させるためである.

3. 予備実験の結果

ここでは, 2. で述べたコンテンツの分類・マッピング手法を Web ディレクトリに適用した予備実験について述べる. 実験では, goo ドメインで提供されているディレクトリサービスから, 「自然科学と技術＞工学」ジャンルに含まれる以下のサブジャンルを対象に, これらサブジャンルに含まれる 157 個の Web コンテンツを手動で選んで実施した(表1参照. ジャンルによっては2階層のものあり). 本実験では, Web コンテンツに含まれるテキスト情報をコンテンツの言語情報とみなして提案手法を適用した. また本実験では, 最下層ジャンルに対応付ける概念ベクトルは, 上記 157 個の Web コンテンツのうちで, それらジャンルに属するものの概念ベクトルを用いて, k-平均法によりクラスタリングしてクラスタの重心ベクトルによって構成した. 重心ベクトルの個数は, ジャンルごとに異なるが, 2から5である. 図4に基準マップの生成結果を, 図5に Web コンテンツの登録結果を示す. 図4, 図5で Ann.mm はジャンルコードが nn.mm のジャンルに対応付けられた概念ベクトルであり, Bnn.mm はジャンルコードが nn.mm のコンテンツを表す. 図4では, 分類情報の階層構造が保存されるかたちで最下層ジャンルに対応付けられた概念ベクトルたちが2次元上に配置されていることが分かる. また, 比較的内容の近いジャンルどうしが2次元上で互いに近い位置に配置されていることが分かる(例えば, 「情報、通信工学」(10.01)と「電気、電子工学＞半導体」(12.03)や, 「環境工学」(03.01)と「自動車工学」(09.01)および「土木工学＞その他」(13.01)など). 一方, 図5では, Web コンテンツが対応するジャンル領域内に配置されていることが分かる. 我々は, 2次元上で互いに近い位置に配置されている Web コンテンツどうしが意味内容的にも互いに類似, 関連していることを, 主観的ではあるが, 確認している.

表1: 実験に用いたジャンル

ジャンルコード	ジャンル名
01.01	医用生体工学
02.01	化学工学
03.01	環境工学
04.01	機械工学＞その他
04.02	機械工学＞ロボット工学
05.01	機能工学
06.01	経営工学
07.01	原子力工学
08.01	材料工学
09.01	自動車工学
10.01	情報、通信工学
11.01	人間工学
12.01	電気、電子工学＞回路
12.02	電気、電子工学＞その他
12.03	電気、電子工学＞半導体
13.01	土木工学＞その他
13.02	土木工学＞ダム、貯水池

4. おわりに

本稿では, 「AssociaGuide」の中核技術として研究開発を進めてきたメタデータを前提とした大量コンテンツの分類・マッピング手法を Web ディレクトリに適用した試みについて述べた. 今後の課題として, 生成した Web コンテンツの2次元マップの有効性に関する定量評価が挙げられる.

参 考 文 献

[1] 藤田悦郎, 宮原伸二, 安部伸治, 林 泰仁:メタデータを用いたコンテンツ空間の可視化手法, 2002 年電子情報通信学会総合大会, D-8-8, 2002.

[2] 藤田悦郎, 宮原伸二, 安部伸治, 林 泰仁:メタデータを用いたコンテンツ空間の可視化手法－概念空間の2次元非線型投影による逐次登録型コンテンツマップの実現－, FIT(情報科学技術フォーラム)2002 一般講演論文集第2分冊, D-41, 2002.

[3] 笠原要, 松澤和光, 石川 勉:国語辞書を利用した日常語の類似性判別, 情報処理学会論文誌, Vol.38, No.7, pp.1272-1284, 1997.

[4] 今野 浩, 山下 浩, 非線形計画法, 日科技連出版社, 1978.

[5] 熊本 睦, 島田茂夫, 加藤恒昭:概念ベースの情報検索への適用－概念ベースを用いた検索の特性評価－, 電子情報通信学会技術報告, AI98-63, 1999.

[6] 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 年電子情報通信学会総合

大会, D-8-7, 2002.

[7] 宮原伸二, 藤田悦郎, 安部伸治, 林　泰仁：散策型コンテンツガイドシステム AssociaGuide, 映像情報メディア学会ヒューマンインフォーメーション研究会発表予定, 2003.

[8] J.W.Sammon,Jr：A Nonlinear Mapping for Data Structure Analysis, IEEE Trans on Computers, Vol.C-18, No.5, pp.401-409, 1969.

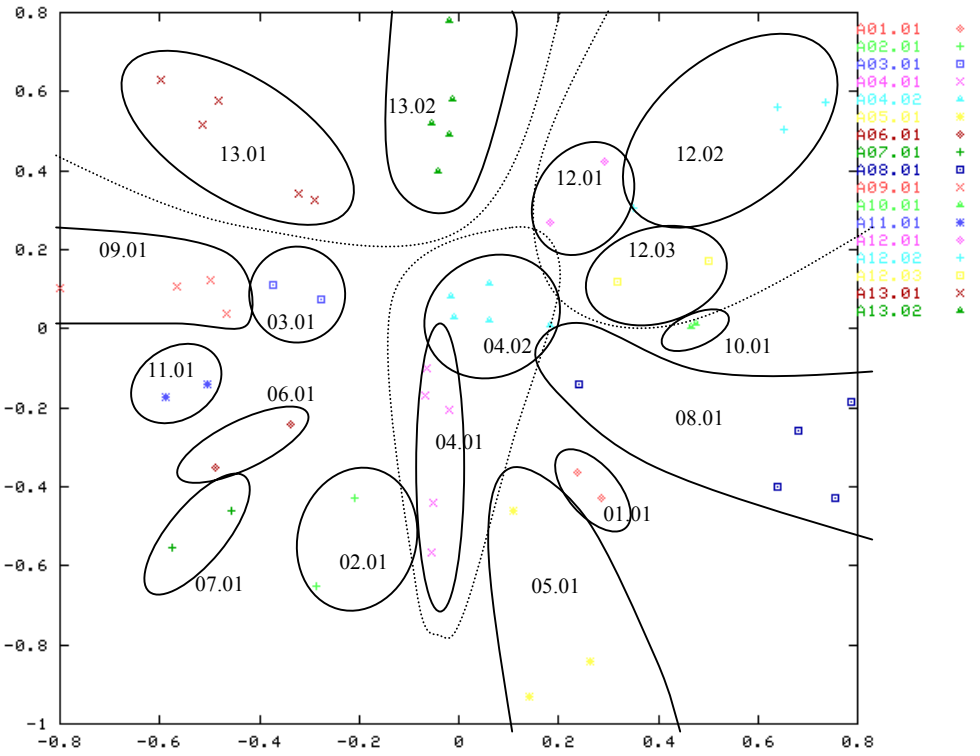

図4：基準マップの生成結果

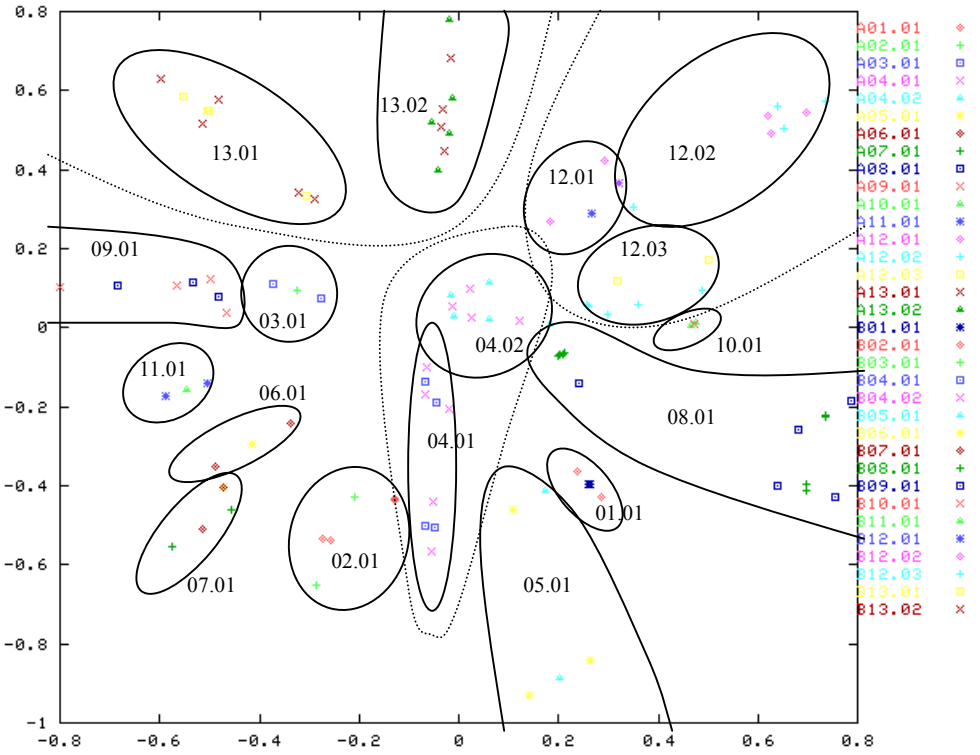

図5：Web コンテンツの登録結果