分類情報と言語情報の統合利用に基づくコンテンツ空間の可視化

Visualization of Content Space based on Classification Information and Textual Information

藤田 悦郎* 宮原 伸二†
安部 伸治† 林 泰仁†
外村 佳伸†

Etsuro FUJITA Shinji MIYAHARA
Shinji ABE Yasuhito HAYASHI
Yoshinobu TONOMURA

本論文では，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・マッピングする新たな手法を提案する．提案手法では，メタデータとしてコンテンツに付与された分類情報と言語情報とを統合的に処理することにより，コンテンツ間の関連性とともに分類情報の情報構造をも反映したコンテンツの２次元マップを生成する．450 件の新聞記事データを用いた予備実験によって提案手法の有効性を確認した．

This paper presents a new framework of 2D visualization of content space based on classification information and textual information. The prominent feature of the proposed method is that content space is mapped into 2D space such that the relevance between content items is reflected based on their classification information and textual information and that those items can be clustered by each classification category of a classification hierarchy to visualize the structure of the hierarchy in the space. The design and implementation of the 2D visualization method are introduced, together with preliminary experimental results.

1．はじめに

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，我々は，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]．本システムでは，ユーザインタフェースにマップ表現を採用することによって，膨大なコンテンツ集合の全体を，電子地図を操作する感覚で概観，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ようにしている．

本システムにおける大量コンテンツの2次元マッピング過程では，コンテンツに付与された，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り，コンテンツ間の意味内容的な関連性とともに，メタデータである分類情報の情報構造をも同時に反映した2次元マップを生成する．この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．これによりユーザは，2次元マップの全体的な構造を容易に把握できるようになるメリットがある．また，コンテンツを探索，散策する際の指標として利用することも可能となる．

一方，上記2次元マッピング過程では，電子地図のメタファーを利用する立場から，新規コンテンツが与えられた場合には，既に登録されているコンテンツの配置は変えずに，新規コンテンツのみを追加配置するようにしている．これによって，ユーザは本システムを使い続ける中で，閲覧したコンテンツをその配置場所に対応づけて記憶したり，逆に配置場所に基づいてコンテンツの内容を推定したりすることができるようになる．いわゆる土地鑑を活かしたコンテンツの探索や散策ができるようになるメリットがある．このような特長は，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などで特に有用であろう．

以下本論文では，本システムで実現しているメタデータを前提としたコンテンツの2次元マッピング手法を中心に検討を進める．まず2章で，追加配置を前提としたコンテンツの2次元マッピング手法について述べる．次いで3章で，予備実験の結果を報告し，4章で関連研究を述べる．最後に5章でまとめと今後の課題を述べる．

2．コンテンツ空間の可視化

提案手法では，まず，分類大系の最下層ジャンルに対応づけられた，ジャンルの意味内容を表す，1つ乃至複数の概念ベクトルを用いて，分類大系を考慮した基準マップを生成する．次に，入力コンテンツが与えられた場合に，メタデータの言語情報から，入力コンテンツの意味内容を表す概念ベクトルを生成して，メタデータの分類情報を考慮しながら基準マップに追加配置する．なお，ここで最下層ジャンルに対応づけた，あるいは入力コンテンツの言語情報から生成した概念ベクトルとは，日本語語彙大系の約 3000 の意味カテゴリへの関連度合いを成分とする多次元ベクトル（概念空間のベクトル）を意味する[2][4][7].

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づけられた概念ベクトルを用いて，メタデータである分類情報の情報構造を反映した基準マップを生成する．すなわち，各々の最下層ジャンルに対応づ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類似性，すなわち，概念空間での距離の近さを考慮しつつ，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考慮して，深さ方向に一致すればするほど，概念ベクトルが2次元上で互いに近い位置に配置されるよう制約を課して概念ベクトルを配置する．これによって，1階層目が同じジャンルに属する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属する概念ベクトルどうしは，その中でもさらに近い位置に配置されることになる．すなわち，最下層ジャンルに対応づけられた概念ベクトルが分類大系の階層構造に同期するかたちで2次元上にクラスタ配置されることになる．一方，上記処理の中では，概念ベクトルどうしの概念空間での距離関係が考慮されるために，分類大系の各階層のジャンルに対応するクラスタ周辺には，意味内容的に近いジャンルのクラスタが配置されることになる．これにより，分類大系の階

* 正会員　西日本電信電話株式会社ソリューション営業本部
e.fujita@bch.west.ntt.co.jp
† 非会員　NTT サイバーソリューション研究所
{miyahara.shinji, abe.sinji, hayashi.yasuhito,tonomura.yoshinobu}@lab.ntt.co.jp

層構造とともに，ジャンルどうしの意味内容的な関連性をも同時に反映した2次元マップが生成されることになる．

以下，基準マップの生成アルゴリズムについて説明するが，ここでは簡単のため，分類大系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容易に拡張可能である．

1階層目のジャンルを $G_p(p=1,\cdots,N_{ROOT})$ とし，G_p 下の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 下の2階層目のジャンル数を表す．2階層目のジャンル(最下層ジャンル) G_{pq} に対応づけられた概念ベクトルの集合を S_{pq} とする．また，1階層目のジャンル G_p に対して S_p を次のように定める．また，S を次のように定める．

$$S_p = S_{p1} \cup \cdots \cup S_{pN_p} \quad (1)$$

$$S = S_1 \cup \cdots \cup S_{N_{ROOT}} \quad (2)$$

S は，すべての最下層ジャンル G_{pq} に対応づ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できるだけ保存されるように，多次元尺度法[9]を用いて2次元上に配置するが，その際，前述した分類大系の階層構造を2次元上に可視化するために，上記概念ベクトルの2次元配置に関してある制約を課す．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．

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

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，分類大系の1階層目の情報構造を可視化するために，$S_p(p=1,\cdots,N_{ROOT})$ 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課す．

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

ここで，n は S に含まれる概念ベクトルの総数 $\#(S)$ を表す． また，d_{S_p,S_q} は S_p あるいは $S_{p'}$ に含まれる概念ベクトルに対応する2次元座標たちの重心を $(\bar{x}_{S_p}, \bar{y}_{S_p})$ 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，

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

とするとき，

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

また，σ_{S_p} あるいは $\sigma_{S_{p'}}$ は，S_p あるいは $S_{p'}$ に含まれる概念ベクトルに対応する2次元座標たちの上記重心を中心とする2乗平均の平方根を表す．すなわち，

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

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 に含まれる概念ベクトルに対応する2次元座標たちによって作られる2次元上のクラスタと，$S_{p'}$ に含まれる概念ベクトルに対応する2次元座標たちによって作られるクラスタとが互いに分離して配置されるようにするための制約である．

同様に，分類大系の2階層目の情報構造を可視化するために，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，次の不等式制約を課す．

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

ここで，$d_{S_{pq},S_{pq'}}$，$\mu_{S_{pq},S_{pq'}}$ および $\sigma_{S_{pq}}$，$\sigma_{S_{pq'}}$ は1階層目の場合と同様にして定義される．式(8)の意味は式(4)と同様であり，S_{pq} および $S_{pq'}$ に対応する2次元上のクラスタが互いに分離して配置されるようにするための制約である．なお式(8)は，すべての $p=1,\cdots,N_{ROOT}$ にわたって導入される．

式(4)，(8)の制約条件を同時に満たす2次元座標の組であって，しかも式(3)を最小化する，すねわち，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，前述の性質を備えた基準マップを生成する．なお，上記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[6].

2.2 コンテンツの登録

ここでは，入力コンテンツにメタデータとして付与された分類情報と言語情報を用いて，入力コンテンツを基準マップに追加的に配置する．以下，入力コンテンツに付与されたジャンルが最下層ジャンル 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表す概念ベクトルを生成する[4][7].

(手順2) 式(3)を模した次式によってコンテンツの2次元座標を決定する．

$$E' = \frac{1}{\sum_i d^*_i} \sum_i \frac{(d^*_i - d_i)^2}{(d^*_i + \delta)^2} \quad (9)$$

ここで，d^*_i は，S に含まれる概念ベクトル v_i と，入力コンテンツの概念ベクトルの概念空間でのユークリッド距離を表し，d_i は v_i に対応する2次元座標 (x_i, y_i) と，入力コンテンツに対応する2次元座標 (x, y) のユークリッド距離を表す．また δ は $0<\delta<1$ を満たす実数である(入力コンテンツには依存しない)．ただし，入力コンテンツが最下層ジャンル G_{pq} に属することを2次元上で反映するために，式(9)の最小化で以下の制約を課す．最下層ジャンル G_{pq} に対応づけられた概念ベクトル集合 S_{pq} のうち，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を v_k とするとき，v_k に対応する2次元座標 (x_k, y_k) と，入力コンテンツに対応する2次元座標 (x, y) のユークリッド距離 d_k が，

$$g_i(x,y) = d_i - d_k \geq 0 \quad (10)$$

ここで，d_i は，S に含まれる，v_k とは異なる概念ベクトル v_i に対応する2次元座標 (x_i, y_i) と，上記 (x, y) のユークリッド距離を表す．式(10)は S に含まれる，v_k とは異なるすべての概念ベクトル v_i にわたって導入される．これにより入力コンテンツは，v_k に対応する2次元座標付近，したがって，基準マップ上のジャンル G_{pq} の領域内もしくは境界付近に配置され、しかも S に含まれる概念ベクトルとの概念空間での距離関係が最大限保存されるような位置に配置されることになる．なお，上記制約つき最小化問題は，2.1と同様，逐次2次計画法を用いて解くことができる[6].

3．予備実験の結果

3.1 実験に用いたデータ

本実験では，適用対象として毎日新聞の 30,207 記事データ(1994年発行分)からなるRWCPコーパスを使用した．この記事データには，記事の掲載面を示す掲載面種別コードとともに，各記

事に対して，複数の UDC コード[3]が付与されている．本実験では分類大系の階層の深さを2とするために，1 階層目の分類に掲載面種別コードを，2階層目の分類に UDC コードを用いた．次表に分類大系の構成を示す．

ジャンルコード	ジャンル名	
	第1階層(掲載面)	第2階層(UDC)
01.01	経済	金融･通商
01.02	経済	経営･広告
02.01	家庭	食品･食事
02.02	家庭	医療･健康
03.01	国際	政治
03.02	国際	法律
04.01	社会	金融･通商
04.02	社会	法律
04.03	社会	医療･健康

表 1　本実験における分類大系

Table1　Hierarchy of Classification Categories

上記分類大系の各々の最下層ジャンルに分類された記事データをそれぞれ 50 件ずつ，したがって，全ジャンルで 450 件の記事データを選んで実験を行った．なお，上記記事データには，記事の見出しおよび本文に関するキーワードが付与されており，本実験では，これを記事データの言語情報として用いた．

3.2 基準マップの生成結果

本実験では、各々の最下層ジャンルに対して次の要領で概念ベクトルを対応づけた．最下層ジャンルに含まれる 50 の各々の記事データから，2.2 の入力コンテンツの場合と同様にして概念ベクトルを生成し，k-平均法によってそれらを 20 のクラスタに分割する．そして各々のクラスタの重心ベクトルを求めることにより，ジャンルの概念ベクトルを生成した．したがって，1ジャンルにつき 20，全ジャンルで 180 の概念ベクトルを対応づけた．これらの概念ベクトルを用いて基準マップを生成した結果を図1に示す．

3.3 コンテンツの登録結果

上記生成した基準マップに対して 450 件の記事データを登録した結果を図2に示す．なお図1および図2で，点線は1階層目のジャンルに対応する領域を，実線は2階層目のジャンルに対応する領域を表している．

3.4 考察

図1では，最下層ジャンルに対応づけられた概念ベクトルが，分類大系の階層構造に同期するかたちでクラスタ配置されていることが分かる．また図1で，ジャンル「社会＞金融･通商」(04.01)の概念ベクトルの存在領域の左下方には，第 1 階層が同じ，ジャンル「社会＞法律」(04.02)のそれが隣接しているが，右下方には，第 1 階層が異なる，ジャンル「国際＞法律」(03.02)のそれが隣接していることが分かる．これは詳細な評価が今後必要ではあるが，提案手法が，最下層ジャンルに対応づけられた概念ベクトルを分類大系に同期するかたちでクラスタ配置しながらも，内容的に近いあるいは連想的に関連するジャンルの概念ベクトルを2次元上で互いに近い位置に配置したためと考えられる．一方，図2では，最下層ジャンルに分類された記事データが，対応するジャンルの領域内もしくは境界付近に配置されており，制約つき最適化問題として定式化した登録アルゴリズムの有効性が分かる．なお，比較のため，ジャンルの概念ベクトルおよび記事データの概念ベクトルを式(3)のみによって，すなわち，従来の多次元尺度法によって2次元配置した結果を図3および図4に示す．これらの結果では，ジャンル「家庭＞食品･食事」(02.01)など一部のジャンルで，概念ベクトルのクラスタ配置が認められるものの，必ずしもすべてのジャンルではクラスタ配置が認められない．これは，従来の多次元尺度法がそもそも概念ベクトルどうしの概念空間での距離関係のみを考慮するものであって，提案手法のように，概念ベクトルとは別に存在する分類情報をも考慮する手法ではないことを考えれば，自然な結果であると言えよう．

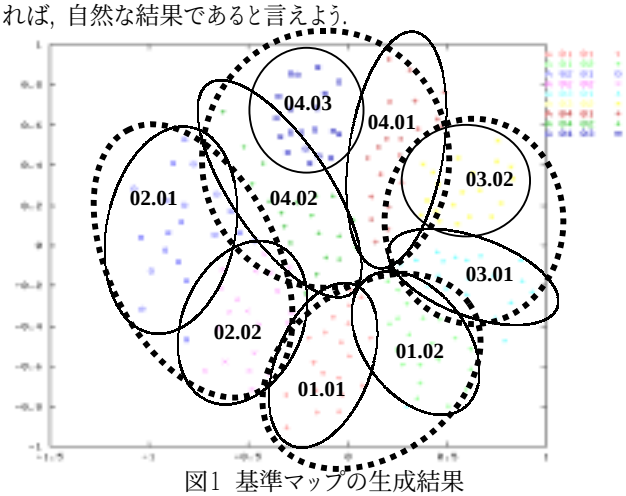

図1　基準マップの生成結果

Fig.1 Generating Result of Basis Map

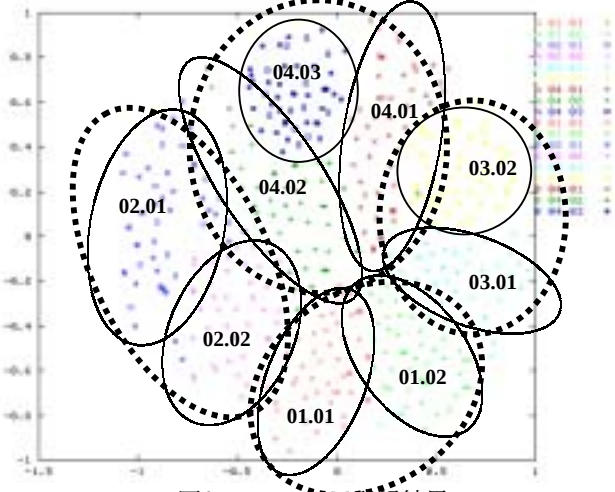

図2　コンテンツの登録結果

Fig.2 Mapping Result of Content Items

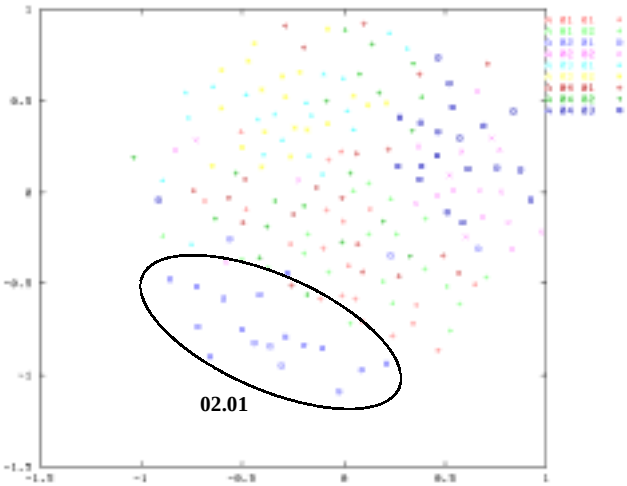

図3　従来法による基準マップの生成結果

Fig.3 Generating Result of Basis Map by Ordinary MDS

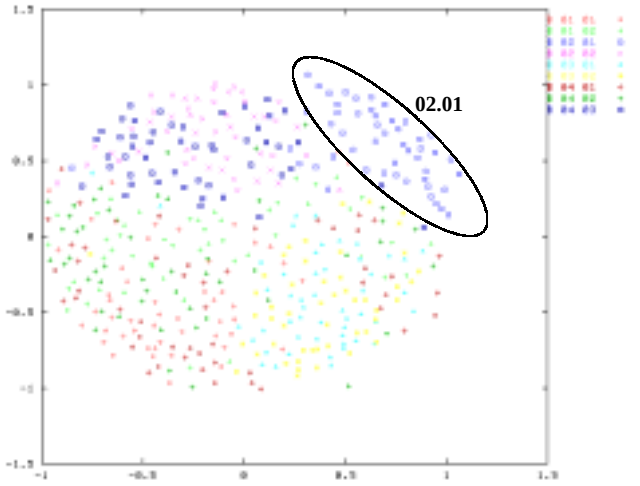

図4 従来法によるコンテンツの2次元配置結果
Fig.4 Mapping Result of Content Items by Ordinary MDS

4. 関連研究

大量文書などの可視化を目的に, 多次元データを2次元上に分類･マッピングする手法はこれまでにも多数提案されている. 例えば, Wise らあるいは舘村らは多次元尺度法あるいはスプリングモデルを用いて大量文書を2次元上に配置する手法を提案している[11][10]. これらは, 文書から生成した特徴ベクトルを用いて, 特徴ベクトルが近ければ近いほど, 2次元上でもそれらを近くに配置しようとするものである. しかしながら, 提案手法のように文書情報とは別に存在する分類情報をも大量文書の可視化に反映しようとするものではない. 一方, 特徴空間における文書のクラスタ構造などマクロな構造を2次元上に可視化する手法が, 例えば吉岡らによって提案されている[12]. 吉岡らは k-平均法と人工神経回路網を組み合せた手法によって, 特徴空間のクラスタ構造を2次元可視化する技術を実現している. クラスタ構造が2次元上で巨視的に表現される点で吉岡らの手法は我々の手法と類似しているが, 吉岡らの手法は上記同様, そもそも分類情報の情報構造を2次元可視化に反映しようとするものではないことから, 我々の手法とは本質的に異なる. 他方, 自己組織化マップ SOM を用いた大量文書の可視化技術がこれまでに多数提案されているが[5][8], SOM はそもそも特徴空間のクラスタを2次元上のセルに投影する手法であって, 提案手法のように文書毎に2次元座標を決定することを前提とした技術ではない. さらに上記同様, 分類情報の情報構造までをも可視化に反映しようとするものではないことから, 提案手法とは本質的に異なる.

5. まとめと今後の課題

本論文では, コンテンツに付与された分類情報と言語情報とを統合的に処理することにより, コンテンツ間の内容的な関連性とともに, 分類大系の階層構造など, 分類情報の情報構造をも反映した2次元マップを生成する方法について提案した. 提案手法は, マップ上に分類情報の情報構造が巨視的に表現されることから, ユーザにとっては, 2次元マップの全体的な構造を容易に把握できるメリットがある. また, コンテンツの追加配置を実現したために, ユーザは使い続けていく過程でいわゆる土地鑑を活かしたコンテンツの探索や散策ができるようになるメリットがある.

今後の課題としては, 提案手法の有効性に関する定量的評価が挙げられる.

[謝辞]

毎日新聞 94 年版に関して, 記事データの研究利用許諾をいただいた毎日新聞社に感謝いたします.

[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－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, D-41, 2002.

[2] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦：日本語語彙大系, 岩波書店, 1999.

[3] 情報科学技術協会：国際十進分類法, 日本語中間版第 3 版, 丸善, 1994.

[4] 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.

[5] T.Kohonen：The Self-Organizing Map, Proc. IEEE, Vol.78, No.9, pp.1464-1480, 1990.

[6] 今野浩, 山下浩, 非線形計画法, 日科技連出版社, 1978.

[7] 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.

[8] 仁木和久, 田中克巳：ニューラルネットワーク技術の情報検索への適用, 人工知能学会誌, Vol.10, No.1, pp.45-51, 1995.

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

[10] 舘村純一：DocSpace：文献空間のインタラクティブ視覚化, インタラクティブシステムとソフトウェアIV 日本ソフトウェア科学会 WISS'96, pp.11-20, 近代科学社, 1996.

[11] J.A.Wise et.al：Visualizing the non-visual: Spatial analysis and interaction with information from text documents, Proc. of IEEE Information Visualization '95, 1995.

[12] 吉岡琢, 高岡善朗, 石井信, 伊東実：WWW 上の文書集合の可視化による検索支援, データベースと Web 情報システムに関する合同シンポジウム(DBWeb2000), pp.143-148, 2000.

藤田 悦郎 Etsuro FUJITA
NTT サイバーソリューション研究所勤務を経て, 現在西日本電信電話株式会社ソリューション営業本部勤務. 1996 京都大学大学院修士課程修了. NTT サイバーソリューション研究所在籍中は, コンテンツガイドシステムの研究･開発に従事.

宮原 伸二 Shinji MIYAHARA
NTT サイバーソリューション研究所勤務. 1999 大阪大学大学院修士課程修了. コンテンツガイドシステムの研究･開発に従事.

安部 伸治 Shinji ABE
NTT サイバーソリューション研究所主任研究員. 1986 北海道大学大学院修士課程修了. コンテンツガイドシステムの研究･開発に従事.

林 泰仁 Yasuhito HAYASHI
NTT サイバーソリューション研究所主幹研究員. 1984 慶応義塾大学大学院修士課程修了. コンテンツガイドシステムの研究･開発に従事.

外村 佳伸 Yoshinobu TONOMURA
NTT サイバーソリューション研究所プロジェクトマネージャ. 1981 京都大学大学院修士課程修了. NTT サイバーソリューション研究所インテリジェントメディアプロジェクトマネージャとしてコンテンツサービスに関する研究･開発に従事.