D-41

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—
Visualization of Content Space based on Metadata

藤田 悦郎† 宮原 伸二† 安部 伸治† 林 泰仁†
Etsuro Fujita Shinji Miyahara Shinji Abe Yasuhito Hayashi

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][2]。本稿では、その中核技術として先に提案した大量コンテンツの 2 次元マッピング手法の改良について報告する。

２．従来のマッピング手法とその問題点

文献[2]で述べたマッピング手法では、大量のコンテンツを、各々のコンテンツに概要を表すテキスト情報と、分類を表すジャンル情報がメタデータとして付与されていることを前提として、以下の手順により 2 次元上に一括で配置する。

（手順１）コンテンツに付随するテキスト情報を概念ベース[3][4]にかけ、コンテンツの意味的特徴を表す概念ベクトルを生成する。概念ベクトルは、日本語語彙体系における約3000の意味カテゴリーに対する関連度をその成分として持つ多次元ベクトルである。

（手順２）上記概念ベクトルを用いてコンテンツに付随するジャンル情報を考慮したコンテンツ間距離を算出する。すなわち、コンテンツ i および j の距離 d^*_{ij} を次式により算出する。

$$d^*_{ij} = w_{pq} \times e_{ij} \quad \cdots\cdots(1)$$

ここで、e_{ij} はコンテンツ i および j の上記概念ベクトルを用いたユークリッド距離であり、w_{pq} は i および j が属するジャンル g_p および g_q に関するジャンル非親和度である。ただし、w_{pq} は $p=q$ のとき（ジャンルが同じとき）は１より小さくし、$p \neq q$ のとき（ジャンルが異なるとき）は１以上にする。これによって、同一ジャンル内での「引力」あるいは異なるジャンル間での「斥力」として作用する。

（手順３）上記距離行列 d^*_{ij} に多次元尺度法[5]を適用してコンテンツを 2 次元上に一括配置する。すなわち、次式を最小化ならしめる 2 次元座標の組によってコンテンツの 2 次元配置を決定する。

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

ここで、d_{ij} はコンテンツ i および j の 2 次元座標を用いたユークリッド距離である。上記最適化によって、コンテンツは概念空間上での d^*_{ij} の意味での距離関係を保存するよう 2 次元上に配置されることになる。

以上の手順によって文献[2]では、意味内容的に近いコンテンツを 2 次元上互いに近い位置に配置するとともにコンテンツをジャンル毎にクラスタ化して配置している。さらに内容的に近いジャンルのクラスタを互いに近い位置に配置してコンテンツ単位の意味的連続性のみならずジャンル単位の意味的連続性をも備えたマップを生成している。

我々は、このようなコンテンツマップは、ユーザの連続的なコンテンツ探索や散策を支援するブラウジングインタフェースとして有用であると考えているが、従来のマッピング手法では、一旦マップを生成してしまうと新たなコンテンツが追加で登録できないという問題点があった。追加登録は、インターネットなどのコンテンツが随時更新されるサービスにおいて必要不可欠な機能である。

３．マッピング手法の改良

提案手法では、コンテンツの分類大系すなわちジャンル構造を 2 次元上に可視化したマップ（以下、基準マップと呼ぶ）を生成し、これにコンテンツを一つずつ登録していく。提案手法は、分類大系の最下層ジャンルに付与した概念ベクトルを用いて基準マップを生成する過程と、コンテンツに付随するメタデータを用いてコンテンツを基準マップに逐次的に登録する過程とからなる。

３．１． 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付与された概念ベクトルに対し、２章の手順２および手順３の処理を適用して、ジャンル単位およびジャンルに付与された概念ベクトル単位の両粒度において意味的連続性を有するマップを生成する。なお、ジャンルに付与する概念ベクトルは複数あってもよい。

３．２． コンテンツの登録

ここでは、コンテンツに付随するテキスト情報とジャンル情報を用いて入力コンテンツを基準マップに登録する。以下では、入力コンテンツの属するジャンルが g_p としてアルゴリズムの概要を述べる。

（手順１）２章の手順１と同様にして、入力コンテンツに付随するテキスト情報から概念ベクトルを生成する。

（手順２）２章の手順２と同様にして、入力コンテンツと各々のジャンルの概念ベクトル k との距離 d^*_k を次式により算出する。

$$d^*_k = w_{pq} \times e_k \quad \cdots\cdots(3)$$

ここで、e_k は入力コンテンツの概念ベクトルと、上記概念ベクトル k とのユークリッド距離であり、w_{pq} はジャンル g_p と概念ベクトル k のジャンル g_q に関するジャンル非親

† NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories

和度である。さらに、ジャンル g_p の概念ベクトルであって上記 d^*_k を最小化ならしめる概念ベクトル k0 を求める。

（手順３） 上記概念ベクトル k の基準マップにおける 2 次元座標を x_k とする。また、時刻 t において x_{k0} を含む近傍領域を $N_{k0}(t)$（$N_{k0}(t)$ は $t=0$ のときは全ての x_k が含まれるよう十分大きくとり、時間とともに大きさを単調に減少させる）、入力コンテンツの 2 次元座標を $x(t)$ とする（$x(0)$ は適当に初期化する）。そして、全ての $x_k \in N_{k0}(t)$ を用いて入力コンテンツの 2 次元座標 $x(t)$ を次式により修正していく。

$$x'(t) = x(t) + \alpha(t)h(d^*_k)[x_k - x(t)] \cdots\cdots\cdots\cdots (4a)$$

$$x(t) = x'(t) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (4b)$$

全ての $x_k \in N_{k0}(t)$ による修正後の $x(t)$ を $x(t+1)$ とする。なお、$\alpha(t)$ は時間とともに単調に減少する正値関数であって、$h(\cdot)$ は時間に依存しない正値の単調減少関数である。

（手順４） $N_{k0}(t)$ と $\alpha(t)$ を徐々に小さくしながら、(4a)および(4b)の修正処理を繰り返す。

以上のアルゴリズムは、自己組織化マップの学習アルゴリズム[6]を参考にしたものであるが、これにより入力コンテンツが、基準マップ上の x_{k0} 近傍であって周辺の x_k との概念空間上での d^*_k の意味での距離関係を反映した位置に配置されることになる。すなわち、ジャンル g_p の内部あるいは境界領域であって、概念ベクトル単位の粒度における意味的連続性を考慮した位置に配置されることになる。故に、コンテンツ単位の意味的連続性も実現されることになり、基準マップの性質と併せれば、提案手法においても従来手法と同様なコンテンツマップが実現されることになる。

４．評価実験

「ドラマ」や「娯楽」など 11 のジャンルに(1 階層で)分類されている 312 個の映像コンテンツを用いて提案手法の評価実験を行った。図１に提案手法により基準マップを生成した結果を示す。また、この基準マップに 312 個の映像コンテンツを登録した結果を図２に示す。ただし、図１、図２において L**はジャンルの概念ベクトル、C**は対応するジャンル L**のコンテンツを表す(L01：情報、L02：ドラマ、L03：娯楽、L04：音楽、L05：教養、L06：映画、L07：趣味、L08：スポーツ、L09：ドキュメンタリー、L10：教育、L11：子供向け)。図１では、ジャンルの概念ベクトルがジャンル毎にクラスタを形成するとともに意味的に近いジャンルが互いに近い位置に配置されていることが確認できる。また、図２では、各々のジャンルのクラスタ内に対応するコンテンツが配置されていることが確認できる。

５．おわりに

本稿では、先に提案した大量コンテンツのマッピング手法の改良技術としてコンテンツの逐次登録を前提とした手法を提案し、評価実験を行って有効性を確認した。今後の課題としては、実験結果を「AssociaGuide」においてユーザの立場から評価することが挙げられる。また、これと関連して基準マップの自動生成の観点から次のような課題が挙げられる。

（１）ジャンルに付与する概念ベクトルの最適化：現在は手動により作成しているが、今後は機械学習技術などを用いて最適な概念ベクトルを自動的に決定する。

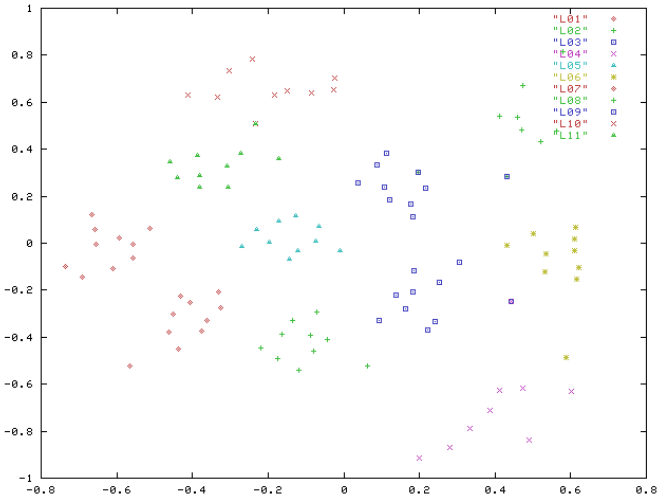

図１ 基準マップの生成結果

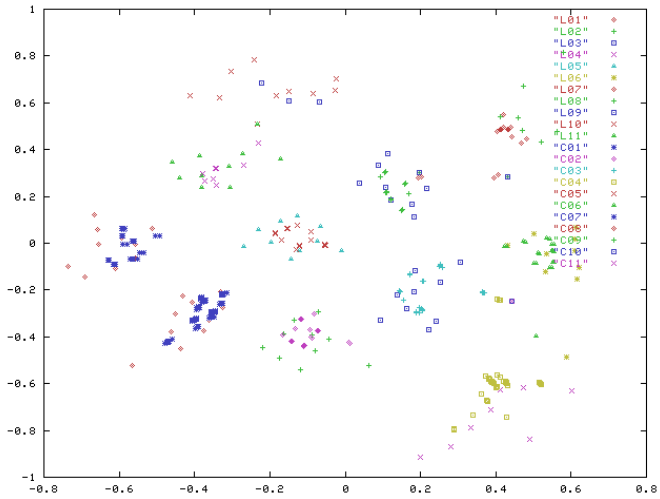

図２ コンテンツの登録結果

（２）ジャンル非親和度 w_{pq} の最適化：現在は手動により設定しているが、今後はコンテンツ探索や散策のしやすさの観点からこれを最適化するアルゴリズムを開発する。

また、ジャンル規模を拡大した場合およびジャンル構造を多階層化した場合の検証やコンテンツ規模を拡大した場合の検証なども併せて進めていく必要があると考えている。

参考文献

[1]宮原他：散策型映像ポータルシステム AssociaGuide の提案, 2002 年信学総大, D-8-7, 2002

[2]藤田他:メタデータを用いたコンテンツ空間の可視化手法,2002 年信学総大, D-8-8, 2002

[3]笠原他：国語辞書を利用した日常語の類似性判別, 情処論, Vol.38, No.7, pp.1272-1284, 1997

[4]熊本他：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－, 信学技報, 人工知能と知識処理, 1999

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

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