

ベクトル空間モデルのデータ分布の全体傾向を考慮した部分一致検索

木 俵 豊^{††} 澤 扶 美^{††} 田 中 克 己^{††}

マルチメディアデータをはじめとして、複数のデータを組み合わせた複合的なデータが幅広く用いられている。このような複合的なデータを作る場合には、それぞれの要素となるデータのふさわしい組合せを考慮する必要がある。本論文では、複数の要素オブジェクトから構成された編集オブジェクトを格納したデータベースから、適切な要素オブジェクトの組合せを検索するベクトル空間モデルを用いた部分一致検索手法を提案した。提案した手法は、決定した一部のオブジェクトとそれに類似したオブジェクトが過去にどのようなオブジェクトと組み合わせられているかを考慮することにより、特異な解を排除することができる部分一致検索手法である。

本研究では、シミュレータによるシミュレーション実験と、3D データの組合せ検索システムによるマルチメディアコンテンツを対象とした実験により、本手法の有効性を検証した。

Partial Match Query Processing using Vector Space Model and Tendency of Data Distribution

YUTAKA KIDAWARA^{†,††} FUMI SAWA^{††} and KATSUMI TANAKA^{††}

We operate composite data that consists of several element data. When we create them, we must consider plausible combinations. In this paper, we describe a partial match query method that considers with combination tendencies of element objects. We define a feature vector of element object to recognize similarity between other element objects. A feature vector of an edited object is concatenation of element object feature vectors. Also, we need to retrieve element objects to combine with other element objects. Therefore we develop a way to process partial match queries considering combination tendency of edited objects. Furthermore, we developed simulator and 3D digital asset retrieve system to confirm useful features of our proposed method.

1. はじめに

従来、テキストや数値データを対象としたデータベースが広く用いられ、単純なマッチングによる検索が行われてきた。しかし現在では、コンピュータの進化により、大量な計算が実時間で可能となり、より複雑な検索が実現できるようになっている。また、その一方で様々なデータがデジタル化されデータベースに格納されており、その上での有効な検索手法の研究が数多くなされている。

マルチメディアデータの検索には、特徴ベクトルなどをを用いた類似検索の手法が多く開発されている³⁾⁷⁾。

これらは、指定されたサンプルデータの特徴ベクトルに、特徴ベクトル空間上で最近傍のものを検索する方式で検索されるのが一般的である。特徴ベクトルは、文書データ中のキーワードの出現頻度情報を基に作成されたり、画像データの色情報等で作られる⁴⁾。通常、これらの特徴ベクトルは高次元なベクトルとなり、計算量も非常に多くなる。そのため、特徴ベクトルを低次元化する事で検索を効率化する方式が開発されている⁵⁾。

一方、一般にマルチメディアコンテンツと呼ばれるマルチメディアデータは、複数のマルチメディアデータを編集し、組み合わせで制作される。このような複数のデータが存在し、それを複合化させたデータを管理する場合には、オブジェクト指向データモデルが用いられるのが一般的となっている。本論文においては、一般的な複合データ作成時における部分一致検索を議論するために、組合せの要素となるデータを要素オ

[†] 通信・放送機構 神戸リサーチセンター

Kobe Research Center, Telecommunications Advancement Organization of Japan

^{††} 神戸大学 大学院自然科学研究科

Graduate School of Science and Technology, Kobe University

ブジェクト, それらを組み合わせて編集した複合的なデータを編集オブジェクトと呼ぶ。

要素オブジェクトを組み合わせて作成される編集オブジェクトを制作する現場においては, 個々のオブジェクトの類似検索(最近傍検索)だけではなく, 編集オブジェクトを構成する要素オブジェクトの組み合わせ自身を検索したいという要求が存在する。

これまでの特徴ベクトルによる検索手法は, 問い合わせベクトルと検索対象となるデータの特徴ベクトルの次元が同じでなければならなかった。しかし, 複合的なデータの検索においては, 部分的な要素を決定し, それらの決定した要素データにふさわしい組合せとなる要素データの検索が必要な場合がある。

複数の要素オブジェクトが組み合わされた編集オブジェクトは, 要素オブジェクトの組合せ方で全く異なる印象を与える。ある特定のオブジェクトを指定して, そのオブジェクト, もしくは, そのオブジェクトに類似したオブジェクトを含むような編集オブジェクトを検索する問題は, 通常は, 部分一致検索(Partial Match Retrieval)の問題として扱われてきている。本論文では, ベクトル空間モデルを前提として, オブジェクトの類似性と, 要素オブジェクトの組み合わせのデータベースの全体的な傾向の両者を考慮に入れた部分一致検索の問題を取り扱う。

このような問題の1つの応用例として, 我々は以前に, 仮想スタジオの素材の組み合わせ検索のための検索手法²⁾を提案している。本論文では, この検索手法を, ベクトル空間モデルのデータ分布の全体傾向を考慮した部分一致検索と位置づけ, 先に提案した検索アルゴリズムを一般化して示すとともに, シミュレーションによる評価実験を行った結果を示す。また, この検索アルゴリズムは, 基本的に, ベクトル空間上に分布するすべてのデータの総当たりを前提としているため, 処理時間がかかるという欠点がある。そこで, 本論文では, 発見的な近似アルゴリズムを提案し, その有効性を検証するために行った評価実験の結果についても述べる。さらに, その有効性を検証するために, 3D素材を対象とした部分一致検索システムを開発し, 提案手法を実際のマルチメディアデータの検索に適用し, その有効性を検証した。

2. オブジェクトの特徴ベクトル

オブジェクト指向データモデルにおいては, データの特徴を示したクラスと, そのクラス構造の変数が異なるインスタンスが存在する。本論文では, あるクラ

スのインスタンスをオブジェクトと呼び, その同一クラスのオブジェクト集合をカテゴリと呼ぶ。

編集オブジェクトは, 複数の要素となるオブジェクトを組み合わせることによって作成される。このような編集オブジェクトは, 要素オブジェクトの組合せ方に, その制作者の意図や個性が表れる。

例えば, 組み合わせるカテゴリを決定して編集オブジェクトを作成したとしても, 制作者によって, カテゴリ内から選択する要素オブジェクトは異なるため, その組合せで作成される編集オブジェクトは大きく異なる。また, 経験の浅い制作者は, 適切な組合せができずに, 奇妙な編集オブジェクトを作成する事も考えられる。従って, 経験の浅い制作者は経験の深い制作者が行った要素オブジェクトの組合せ傾向を参考にすることで, より適切な構成を持つ編集オブジェクトの作成を行うことができる。

従って, 複合化したオブジェクトを効率よく作成するためには, 要素オブジェクトのデータベースを整備する一方で, 過去に要素オブジェクトを組み合わせて制作した編集オブジェクトのデータベース化を行い, 類似したオブジェクトの検索や過去の傾向を反映した要素オブジェクトの組合せの探索などが必要となる。本章では, 各要素オブジェクトや編集オブジェクトに対する特徴ベクトルの定義について述べる。

2.1 要素オブジェクトの特徴ベクトル

特徴ベクトルを用いた検索手法においては, その作成方法が重要となる。特徴ベクトルを作成する場合には, ベクトルの各要素がそれぞれに影響を及ぼすことのない, 独立した要素を選択する必要がある。しかし, 画像や音声などのマルチメディアコンテンツの場合には, 特徴を記述する際に印象語等が用いられ, その言葉の曖昧性により各印象語の独立性を判断することは容易ではない。このような問題に対して意味の数学モデルによる特徴ベクトルの生成などが研究されている。⁸⁾ 本提案手法で印象語を用いた特徴ベクトルを作成する場合においては, これらの手法を用いて直交性を保証した特徴ベクトルを作成する必要がある。

本論文では, ベクトル空間モデルを用いた部分一致検索手法について議論を中心とし, 各要素オブジェクトへ付与される特徴ベクトルは各要素の独立性が保証されているという前提で議論する。

従って, 要素オブジェクトの特徴を表現する直交性を持った m 個の特徴量 $f_i (0 \leq i \leq m)$ を持つ要素オブジェクト o の特徴ベクトル $vector(o)$ を, それぞれの特徴の程度を数値化したベクトルとして, 以下の様

に定義する

$$\text{vector}(o) = (f_1, f_2, \dots, f_m)$$

この特徴ベクトルは、各オブジェクトの特徴を表すもので、独立性があいまいな要素を含まないものとする。本論文における実験においては、HSV 値、部品数などのオブジェクトの特徴を表現し、各要素間に関連性を持たない独立したものを用いている。

2.2 編集オブジェクトの特徴ベクトル

要素オブジェクトを組み合わせで作成した編集オブジェクトは、多数の要素から構成される複合的なデータである。先に定義したように各要素オブジェクトはあるカテゴリを持っており、そのカテゴリ内の異なるオブジェクトを選択することで、同一の構造であるが異なる特徴を持つ編集オブジェクトの構築が可能である。

従って、 n 種類のオブジェクトを組み合わせた編集オブジェクト o_{edit} は、以下の様に定義する。

$$\begin{aligned} o_{edit} &= (o_{c_1}, o_{c_2}, \dots, o_{c_n}) \\ o_{c_1} &\in C_1, o_{c_2} \in C_2, o_{c_n} \in C_n \end{aligned}$$

編集オブジェクトの特徴は、組み合わせられる要素オブジェクトのカテゴリと、そのカテゴリに含まれ、選択された構成要素の特徴で表現できる。従って、編集オブジェクトの構成要素となる要素オブジェクトの特徴ベクトルを用いて編集オブジェクトの特徴ベクトルを表す事が可能である。編集オブジェクトは、それぞれ独立した要素オブジェクトから構成されているので、連結された特徴ベクトルの各要素も独立しており、直交性を有する。従って、各要素オブジェクトの特徴ベクトルの直積によって得られたこのベクトルを、編集オブジェクトの特徴ベクトルと定義する。

従って、 n 個の要素オブジェクトを組み合わせた編集オブジェクト o_{edit} の特徴を表現する特徴ベクトル $\text{vector}(o_{edit})$ は、要素オブジェクトの特徴ベクトルの直積で表され、要素オブジェクトの特徴ベクトルを連結させたベクトルとして、以下の様に定義する。

$$\begin{aligned} \text{vector}(o_{edit}) &= \text{vector}(o_1) \times \text{vector}(o_2) \times \dots \times \text{vector}(o_n) \\ &= (f_{o_{c_1}1}, f_{o_{c_1}2}, \dots, f_{o_{c_1}m}, f_{o_{c_2}1}, \dots, f_{o_{c_n}m}) \end{aligned}$$

但し、 $\times$ はベクトルの直積を表す。

3. ベクトル空間モデルを用いたオブジェクトの検索

3.1 オブジェクトの類似度

特徴ベクトル空間における類似度検索は、文書データベースや画像データベースにおいて広く利用されている。データベースに格納されたデータは、キーワードの出現頻度や色情報などを用いて特徴ベクトルを作成する。また、問い合わせもベクトルとして表現され、与えられた問い合わせベクトルに対して、ベクトル間の距離やコサイン相関値等を用いて類似度を特定し、その類似度が最大のオブジェクトを選択する。

我々が提案する手法においてもオブジェクトの類似度は距離を用いた評価関数で表す。これまでに述べたように、要素オブジェクト、編集オブジェクトには特徴ベクトルが付加されている。この特徴ベクトルを用いて類似度が判定される。

例えば、 m 次元の特徴ベクトルを持つ要素オブジェクト a, b があるとする、特徴ベクトルはそれぞれ、

$$\begin{aligned} \text{vector}(a) &= (f_{a1}, f_{a2}, \dots, f_{am}) \\ \text{vector}(b) &= (f_{b1}, f_{b2}, \dots, f_{bm}) \end{aligned}$$

として表される。特徴ベクトルによる類似度は、ベクトル間距離やコサイン相関値によって計算される。本研究においてはベクトル間距離を用いた類似度を $\text{sim}(a, b)$ として定義する。そして、ベクトル間距離を $\text{distance}(a, b)$ として定義する。比較するベクトル a, b の各要素が全て一致した場合にベクトル間距離が 0 となり、類似度 $\text{sim}(a, b)$ は最大値として正の実数値 s_{max} となるとする。そして、ベクトル間距離が大きくなるにつれて比較するオブジェクトは似ていないものとなるので類似度は低下する。これらを表現するために、類似度計量関数 $\text{sim}(a, b)$ を距離を変数とする任意の単調減少関数 $G(\text{distance}(a, b))$ によって定義する。従って、ベクトル間距離と類似度を以下の様に定義する。

$$\begin{aligned} \text{distance}(a, b) &= \sqrt{\sum_i (f_{ai} - f_{bi})^2} \\ \text{sim}(a, b) &= G(\text{distance}(a, b)) \end{aligned}$$

但し、

$$G(x) = \begin{cases} s_{max} & x = 0, (s_{max} \text{ は正の実数値}) \\ 0 & x = \infty \end{cases}$$

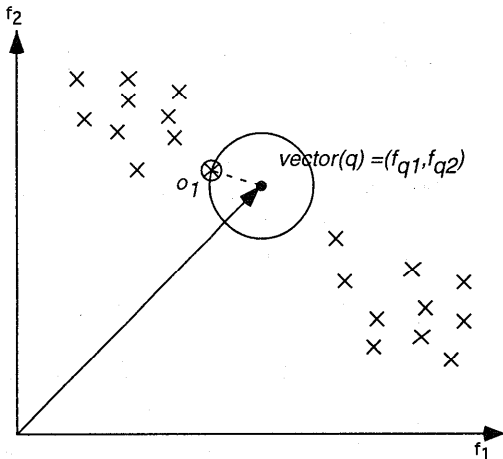


図1 問い合わせベクトルと最近傍解
Fig. 1 Query vector and the nearest answer

$G(x_1) > G(x_2), (x_1 < x_2)$ を満たす単調減少関数とする。

3.2 ベクトル空間における要素オブジェクトの検索

要素オブジェクトに特徴ベクトルを付与することで、そのオブジェクトの特徴を定量化した。これを用いて、ユーザの要求するオブジェクトに類似したオブジェクトを検索することを考える。質問ベクトル $vector(q) = (f_{q1}, f_{q2}, \dots, f_{qn})$ が与えられたとき、データベースに含まれる全てのオブジェクトの類似度を調査し、問い合わせに対して最も近いオブジェクトを選択する。例として2次元の特徴ベクトルを持つオブジェクトの検索を図1に示す。この例では、問い合わせベクトル $vector(q) = (f_{q1}, f_{q2})$ に対して、最も近いオブジェクトである o_1 が選択される。

3.3 ベクトル空間における編集オブジェクトの検索

編集オブジェクトのベクトル空間における検索は、要素オブジェクトの検索と同様に類似度を用いて行われる。あるカテゴリの組合せによって構築された編集オブジェクトは、それぞれのカテゴリに含まれる複数の要素オブジェクトの組合せによって多数の編集オブジェクトが作成される。例えば、3次元仮想空間の構築を行う場合、部屋を表す編集オブジェクトは、壁、机、床、椅子等を組み合わせて制作される。このカテゴリの組合せは部屋という概念を表す物であり、そこで制作された和室、子供部屋などは、部屋オブジェクトとして扱うことができる。従って、 n 種類の異なるカテゴリの要素オブジェクトを組み合わせで構築した

編集オブジェクトを格納したデータベース H を以下の様に定義する。

$$H = \{(o_{c1}, o_{c2}, \dots, o_{cn}) \mid o_{c1} \in C_1, o_{c2} \in C_2, \dots, o_{cn} \in C_n\}$$

このデータベースに対して、ユーザが問い合わせを行うと問い合わせベクトルが作成され、問い合わせベクトルに最も類似した組合せを持つ編集オブジェクトを検索する。この問い合わせ q は、以下の様に表される。

$$q = (o_{q1}, o_{q2}, \dots, o_{qn}) \\ o_{q1} \in C_1, o_{q2} \in C_2, \dots, o_{qn} \in C_n$$

従って、各要素オブジェクトが m 次元の特徴ベクトルを持つとすると、問い合わせ q の特徴ベクトルは、以下の様に表される。

$$vector(q) \\ = vector(o_{q1}) \times vector(o_{q2}) \times \dots \times vector(o_{qn}) \\ = (f_{o_{q1}1}, f_{o_{q1}2}, \dots, f_{o_{q1}m}, f_{o_{q2}1}, \dots, f_{o_{q2}m})$$

この問い合わせベクトルに最も類似度が高い編集オブジェクトが問い合わせの答えとして、ユーザに提供される。図2に例を示す。

この例は、要素オブジェクトが1次元の特徴ベクトルを持ち、2つの要素オブジェクトのカテゴリから組み合わせられる編集オブジェクトの類似検索を示している。そして、問い合わせ $q = (o_{q1}, o_{q2})$ に対して特徴ベクトル間の距離が最も近く、最も類似性の高い編集オブジェクト h_1 が問い合わせの解として選択される事を示している。

3.4 ベクトル空間を用いた検索手法の問題点

これまでにベクトル空間モデルを用いた要素オブジェクト、編集オブジェクトの検索について述べた。これまでの手法は非常に有効であるように見える。しかし、この手法はベクトルの次元が完全に一致している事が前提であったが、編集オブジェクトの検索においては、ユーザがその組合せの全てを指定できない場合が考えられる。例えば、仮想空間の構築を行う場合には、様々な要素オブジェクトを組み合わせで編集オブジェクトが制作される。前述の部屋を構築する場合を考えると、部屋には壁、机、床、椅子等の要素オブジェクトが必要となるが、ユーザによっては全ての要素オブジェクトを決定できずに壁や机だけを決定し、それにふさわしい組合せとなる床や椅子の組合せを検

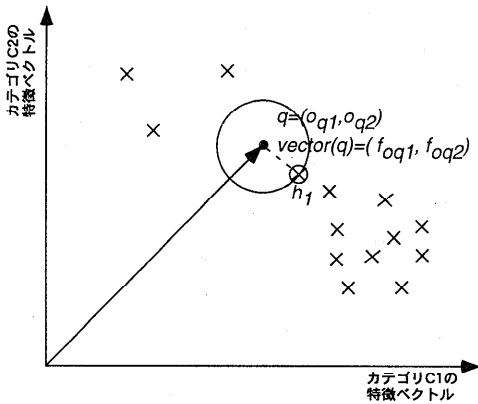


図2 編集オブジェクトの問い合わせベクトルと最近傍解

Fig. 2 Query vector and the nearest answer of edited content

索したいと考えるかもしれない。このような場合には、部分的な問い合わせから、それにふさわしい組合せを検索する部分一致検索が必要がある。しかし、単純な部分一致検索ではユーザの要求に十分な解を得られない場合がある。

例えば、 c_a 、 c_b となる2つのカテゴリの要素オブジェクトを組み合わせた編集オブジェクトが格納されたデータベース H があるとする。またデータベース中には (o_{ca1}, o_{cb1}) 、 (o_{ca2}, o_{cb2}) 、 (o_{ca3}, o_{cb2}) という編集オブジェクトがあるとする。この時、 o_{ca1} 、 o_{ca2} 、 o_{ca3} はそれぞれ類似度が高く、 o_{cb1} 、 o_{cb2} の類似度は低いとする。このような場合に、 o_{ca1} にふさわしい組合せとなる c_b カテゴリの要素オブジェクトを問い合わせる部分一致質問として $(o_{ca1}, *)$ が与えられたとする。この質問に対して、単純な部分一致検索を行うと解は、 o_{cb1} となる。しかし、データベース内の全体の傾向から判断すると、 o_{ca1} に類似した要素オブジェクトは o_{cb2} を選択しており、 o_{ca1} の解としてふさわしい組合せとなるのは、 o_{cb2} となるかもしれない。従って、ベクトル空間モデルを用いた部分一致検索の場合には、単純に決定要素のオブジェクトに最も類似したオブジェクトを選択するのではなく、その他の類似度が高いオブジェクトがどのようなオブジェクトと組み合わせられているか、全体の傾向を判断して、解を決定する事が望ましい。

4. ベクトル空間モデルの組合せ傾向を考慮した部分一致検索

4.1 ベクトル空間による部分一致検索

検索対象となるデータベース H に含まれるオブジェクトが m 次元の特徴ベクトルを持つ物とする。この

時 k 次元の問い合わせベクトル ($k < m$) が与えられると、類似度の判定が不可能となる。類似度を計算するためには、同次元の特徴ベクトルを作成する必要があるため、問い合わせベクトルの決定された要素が座標軸となる次元へ特徴ベクトルを写影させた部分特徴ベクトルを作成し、そのベクトル間の距離を計算し類似度を判定することが必要である。

例えば、 n 個のカテゴリから構成された編集オブジェクトが含まれたデータベース H に対して、問い合わせ q を与えたとする。この時、問い合わせは以下の様になる。

$$q = (o_{c1}, *, o_{c3}, \dots, *, o_{cl}, \dots)$$

但し、 $l < n$ であり、

"*" は検索要求カテゴリを表す。

そして、問い合わせベクトルは以下のように定義される。

$$\Pi_{C_1, \dots, C_l} \text{vector}(q) =$$

$$\text{vector}(o_1) \times \text{vector}(o_3) \times \dots \times \text{vector}(o_l)$$

但し、 $\Pi_{C_1, \dots, C_l}$ は、 $C_1, \dots, C_l$ への写影を表す。

そこで、部分一致検索に用いられる類似度は、データベース中の編集オブジェクト h の特徴ベクトルを問い合わせベクトルの決定要素の次元に写影させたベクトルとの類似度になり、以下のように定義される。

$$\text{sim}(\Pi_{C_1, \dots, C_l} \text{vector}(q), \Pi_{C_1, \dots, C_l} \text{vector}(h)) \\ h \in H$$

図3に例を示す。この例では、簡単のため、2つのカテゴリから組み合わせられる編集オブジェクトに対する検索を行うことを考える。また、組み合わせられる要素オブジェクトは1次元の特徴ベクトルを持つ物とする。そして、編集オブジェクトのデータを格納するデータベースを以下の様に定義する。

$$H = \{(o_{c1}, o_{c2}) \mid o_{c1} \in c_1, o_{c2} \in c_2\}$$

そして、問い合わせ $q = (q_1, *)$ が与えられたとする。この問い合わせは、直観的には " c_1 カテゴリに含まれる q_1 オブジェクトにふさわしい組合せとなる c_2 カテゴリのオブジェクトを検索せよ" というものである。この時、類似度は以下のように表される。

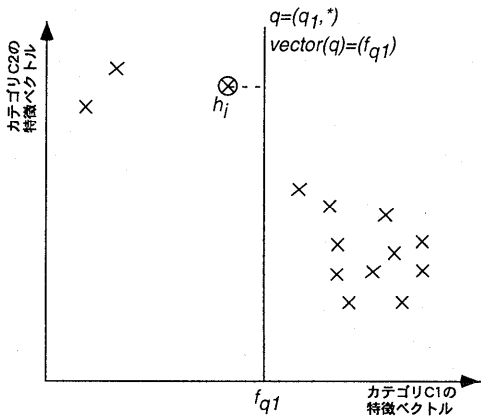


図 3 部分一致検索における最近傍解

Fig. 3 The nearest answer of partial match query processing

$$\text{sim}(\Pi_{c_1} \text{vector}(q), \Pi_{c_1} \text{vector}(h)) \\ h \in H$$

図3の例では、類似度の最も大きい $h_i = (o_{i1}, o_{i2})$ が選択される。そして、 h_i のカテゴリ c_2 の要素 o_{i2} が、 q_1 に最もふさわしい組合せとして選択される。その結果、 (q_1, o_{i2}) が問い合わせ q の解とされる。これは直観的には、 q_1 と組み合わせられた編集オブジェクトは存在しなかったので、 q_1 に最も類似している o_{i1} の組合せを調べて、その組合せ相手であった o_{i2} が最もふさわしい組合せと判断している。

4.2 全体の傾向を考慮したふさわしい組み合わせの検索

先に述べた部分一致検索では直観的には類似度が最も高いふさわしい組合せを選択しているといえる。しかし、それが適切な答えだとは限らない。例えば、図3の様な場合には、解として最も近い h_i が選択されるが、全体の分布を見てみると少し離れたところにデータのクラスタが存在する。問い合わせ q の意味は、 q_1 にふさわしい組合せとなるカテゴリ c_2 の要素オブジェクトを発見する事であり、これらのクラスタの存在を無視できない。もしかすると解として選んだ h_i は特殊な組合せであり、むしろ、多くの組合せが存在するクラスタ内のデータを選択すべきかもしれない。このような編集オブジェクト h_i を選択した原因は、データベース内の組合せ傾向を考慮しておらず、単純に類似度の高い物を選んでいるからである。特にふさわしい組合せを持つ編集オブジェクトを部分的な問い合わせで検索する場合には、全体の傾向を考慮しながら、最もふさわしいと考えられる組合せを選択する必

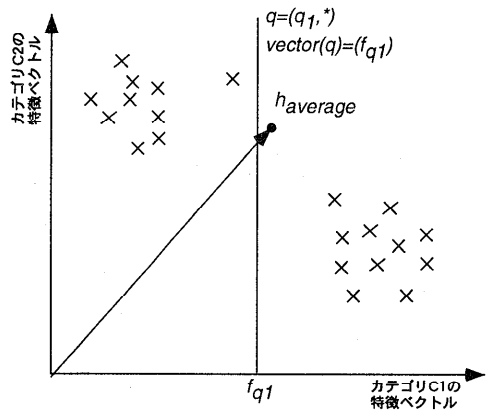


図 4 平均ベクトルによる組合せ傾向の表現

Fig. 4 Combination tendency using an average vector

要がある。

4.3 組合せ傾向の表現

データベース内の組合せ傾向を表現する最も単純な手法は、平均ベクトルを計算することである。しかし、図4の様な複数のクラスタが存在するような場合には、平均ベクトルは、全く特徴を表現できなくなる。これは、全ての特徴ベクトルを同一の尺度で平均化している事に原因がある。

図4の様な部分一致検索の場合にユーザが求めているのは、 q_1 にふさわしい組合せとなるカテゴリ c_2 の要素オブジェクトである。従って、 q_1 や q_1 に類似した物が、カテゴリ c_2 のどのようなオブジェクトと組み合わせられているかを考慮することが大切である。つまり、部分一致質問ベクトルとデータベース内のデータとの類似度が高いものを重視した組合せ傾向を決定する必要がある。そこで、以下のような重み付け平均ベクトルを算出する。

$$h_{w_average} = \frac{\sum_{h \in H} (\text{sim}(\Pi_{c_1, \dots, c_n} \text{vector}(q), \Pi_{c_1, \dots, c_n} \text{vector}(h)) \bullet \text{vector}(h))}{|H|}$$

但し、 $|H|$ はデータベース H に含まれる編集オブジェクトの総数

この重み付け平均ベクトルは問い合わせベクトルに対して、類似度の高い要素を含む組合せを重視した傾向を表すものである。従って、このベクトルと問い合わせベクトルとの交点を最もふさわしい組合せの解と考える。しかし、その点に一致する組合せが存在すると

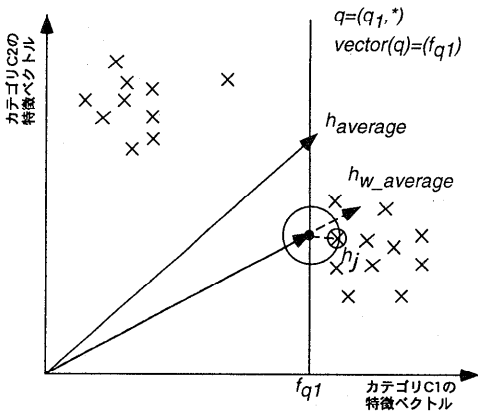


図5 重み付け平均ベクトルによる解の探索

Fig. 5 Search answer using an weight vector

は限らないため、その交点を終点とするベクトルと最も類似度の高い特徴ベクトルを持つ編集オブジェクトを探索し、それを解とする。図5に例を示す。

この例では、問い合わせ $q = (q_1, *)$ に対して、重み付け平均ベクトル $h_{w_average}$ を計算し、問い合わせベクトルとの交点を終点とするベクトルを求めた後に、そのベクトルと特徴ベクトルの類似度が最も高い編集オブジェクトを探索する。そして、探索された編集オブジェクト $h_j = (o_{j1}, o_{j2})$ の要素である o_{j2} を、問い合わせ q に対するふさわしい組合せの解として (q_1, o_{j2}) を得る。

この手法は、問い合わせベクトルに類似した編集オブジェクトを重視した組合せ傾向を考慮して解を求めるため、問い合わせに偶然一致する特異な編集オブジェクトの選択を防ぐことができる。しかし、この手法は、全てのデータに対して総当たり計算をする必要があるが、ユーザによって問い合わせが決定されなければ類似度の計算が不可能なため、対象となる編集オブジェクトや特徴ベクトルの次元が増加すると実時間での検索が不可能となるおそれがある。そのため、高次元な特徴ベクトルを持つ要素オブジェクトや大量のデータを対象とする場合には、計算量を削減した手法が必要となる。

4.4 クラスタリング情報を用いた近似手法

計算量を削減するためには、対象となるデータの数減らす必要がある。そこで、ある一定距離にあるデータをクラスタリングし、その代表点となるセントロイドでそのクラスタに含まれるデータを表現する手

法が考えられる。このようなクラスタリング情報を利用することは、計算量の削減において非常に有効である。たとえば、前述の重み付け平均ベクトルをクラスタのセントロイドと問い合わせベクトルとの類似度を用いて近似した重み付け平均ベクトルを用いれば、計算量を大幅に減少させることができる。しかし、この手法では、クラスタの大きさやクラスタ内のデータ数などが考慮されないため、全体の傾向を反映することが困難である。そこでクラスタのセントロイドを解としたときの重み付け平均ベクトルをあらかじめ計算することで検索時の計算量を減少させる手法を提案する。まず、クラスタのセントロイドを問い合わせベクトルとしたときの、全てのデータとの類似度を計算し、重み付け平均ベクトルを算出する。 i 番目のクラスタのセントロイド c_i を問い合わせベクトルとしたときの重み付ベクトル $h_{c_i_w_average}$ を以下のように定義する。

$$h_{c_i_w_average} = \frac{\sum_j \text{sim}(c_i, h_j) \bullet \text{vector}(h_j)}{|H|}$$

但し、 $|H|$ はデータの総数、 $h_j \in H$ である。

この重み付け平均ベクトルは、対象としたセントロイドに類似した編集オブジェクトを重視した要素オブジェクトの組合せ傾向を表している。また、部分一致質問ベクトルがセントロイドを通る場合には、セントロイドの重み付け平均ベクトルがユーザの問い合わせベクトルに対する重み付け平均ベクトルとなる。従って、セントロイドの重み付け平均ベクトルをあらかじめ計算しておけば、それらを用いて近似する事により、問い合わせ時に部分一致質問ベクトルの重み付け平均ベクトルの計算を削減できる。以下に、問い合わせ時に行われる近似重み付け特徴ベクトルの計算手法について述べる。

まず、問い合わせベクトルがユーザによって決定された時には、セントロイドによって計算された重み付け平均ベクトル $h_{c_i_w_average}$ と問い合わせベクトルとの類似度を調べ、類似度の高いセントロイド重み付け平均ベクトルを用いて、問い合わせに対する近似した重み付け平均ベクトルを計算する。これらの手順を図6、図7を用いて具体的に説明する。

- 2つ以上のセントロイド間に 部分一致質問ベクトルが存在する場合

図6の例では、2つのクラスタが存在し、問い合わせベクトルと同次元のセントロイドを通るベク

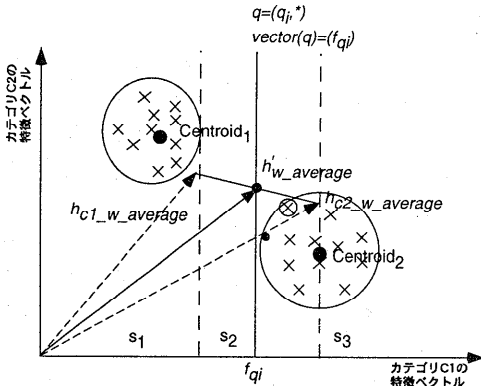


図 6 クラスタリング情報を用いた近似手法 (case 1)
Fig. 6 An approximate method using a centroids of clusters (case 1)

トルによって分割化された3つの空間がある。そしてセントロイドを問い合わせとした時の重み付け平均ベクトル $h_{c1_w_average}$, $h_{c2_w_average}$ があらかじめ計算されているものとする。この時, $centroid_1$ と $centroid_2$ の間に問い合わせベクトルが存在する問い合わせ $q = (q_i, *)$ が与えられたとすると, $h_{c1_w_average}$, $h_{c2_w_average}$ それぞれの終点を結んだベクトルと問い合わせベクトルとの交点を計算し, その交点を近似した重み付け平均ベクトル $h'_{w_average}$ の終点とする。そして, $h'_{w_average}$ に最も類似度の高い編集オブジェクトを解とする。

- セントロイド間に部分一致質問ベクトルが存在しない場合

次に図7の様に, 質問ベクトルがセントロイド間に存在しない問い合わせ $q = (q_j, *)$ が与えられたとする。この場合には, 最も近いクラス, 即ち, $centroid_2$ を含むクラス内のオブジェクトの影響を強く受ける。そこで, 最も近いクラス, 重み付け平均ベクトル $h_{c2_w_average}$ を質問ベクトルに写影させたベクトルを $h'_{w_average}$ として決定し, そのベクトルと類似度の最も高い編集オブジェクトを解として選択する。

5. シミュレーションによる評価

これまでに述べた手法の有効性を検証するために Java 言語で開発したシミュレータを用いてシミュレーションを行った。検索対象は, 1次元の特徴ベクトルを持つ要素オブジェクトを2種類組み合わせた編集オブジェクト (o_x, o_y) である。その編集オブジェクトに

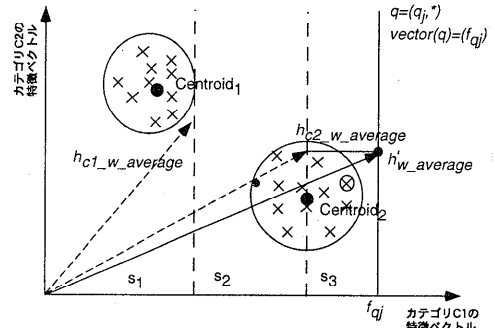


図 7 クラスタリング情報を用いた近似手法 (case 2)
Fig. 7 An approximate method using centroids of clusters (case 2)

対し, 問い合わせ $(q_x, *)$ を与え, 最近傍検索, 提案手法 (総当たり計算), 簡略化提案手法 (近似計算手法) の3種類の解を求め, 最近傍検索手法の解と総当たり提案手法の解との誤差, 簡略化提案手法の解と総当たり提案手法の解との誤差を評価した。

検索対象は, 10個のクラスを形成する1000個のデータである。検索対象となるオブジェクト h は, 以下のような特徴ベクトルを持つものとする。

$$\begin{aligned} \text{vector}(h) &= (f_x, f_y) \\ 0 \leq f_x \leq 20, 0 \leq f_y \leq 20 \end{aligned}$$

実験は, 同一のデータに対して, $q=(q_x, *)$, $0 \leq q_x \leq 20$ となる問い合わせを与える。実験では, q_x を0.1刻みに変化させた200種類の問い合わせベクトルによる検索 (実験1) と, 部分一致質問ベクトルを $(10, *)$ に固定し, 1000パターンの異なるデータに対する検索 (実験2) を行った。全ての実験において, ベクトル間距離が近いものを特に重視した類似度とするために特徴ベクトル a, b 間の類似度計量関数 $\text{sim}(a, b)$ をを以下のように定義した。

$$\begin{aligned} G(x) &= \frac{1}{\alpha \cdot x + 1} \\ \text{sim}(a, b) &= G(\text{distance}(a, b)) \\ &= \frac{1}{\alpha \cdot \sqrt{\sum_i (f_{ai} - f_{bi})^2} + 1} \end{aligned}$$

但し, α は任意の正の実数値

本実験においては, パラメータ α を $\alpha = 5.0$ とした。そして実験1の実験結果を図9, 図10に, 実験2の実験結果を図11に示す。図9の縦軸は, 問い合わせの解として選択された C_2 カテゴリの o_y オブジェクトの1次元特徴ベクトルの値を表している。また, 図10, 図11の縦軸は, 総当たり提案手法で得られた o_y オブジェクトの特徴ベクトルの値を基準として, 簡

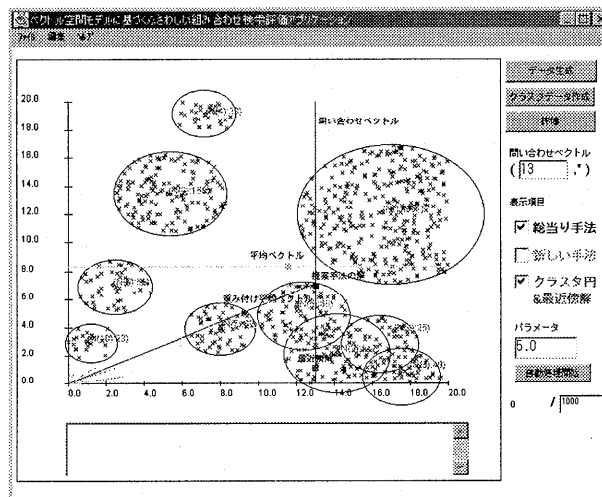


図 8 シミュレータによる実験例

Fig. 8 Simulation example using simulator

略化手法と最近傍検索手法で得られた o_y オブジェクトの特徴ベクトルの値との誤差を示している。

図 9 を見ると最近傍検索の解は、振動を起こしている。これは、全体の傾向を考慮せずに問い合わせに最も近いオブジェクトを選択している事が原因である。それに対して、提案手法と近似手法は、若干異なる部分があるものの、同様な解の傾向を表している。従って、問い合わせベクトルに対して類似性の高い組合せの傾向が反映される提案手法の有効性を示しているといえる。そこで、総当たり計算を行う提案手法の解が最もふさわしい組合せであると仮定し、その他の手法の解について評価する。

図 10 の誤差を見てみると、最近傍検索手法と総当たり提案手法の誤差は、簡略化提案手法と総当たり提案手法の誤差よりも大きい。この傾向は、図 11 の結果でも同様であり、殆どの部分で近似手法が最近傍検索手法の解よりも信頼性が高いといえる。従って、簡略化提案手法においても、単純な最近傍検索手法と比較すると、よりふさわしい組合せとなる解の探索が可能であるといえる。

これらの結果から、全体の組合せ傾向を考慮した提案手法は、従来の部分一致検索における最近傍検索では困難であったふさわしい組合せの検索が実現できることが検証できた。

6. 3D 素材データを用いた実験

本章ではマルチメディアデータとして 3 次元の家具のデータを用いた実験について述べる。

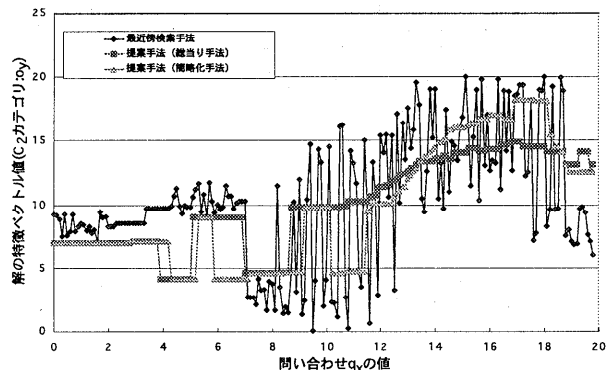


図 9 シミュレーション結果 (3 種類の解の比較)

Fig. 9 Simulation Results (Comparison of three kinds of answers)

6.1 特徴ベクトルの決定と記述

実験にあたってオブジェクトの特徴の記述、組合せの登録、検索を行うためのプロトタイプシステムを開発した。開発したシステムでは、特徴ベクトルの記述、編集オブジェクトの作成と登録、最近傍検索手法による部分一致検索と提案手法による部分一致検索が実行できる。検索結果は、類似度によって順位付けられた候補補リストとして表示される。

本実験ではマルチメディアデータとして机、椅子、その他の 3 個のカテゴリに分類した DXF フォーマットの 3 次元データを合計 100 個用意した。3 D データにおいては、形状はモデリング情報として記述され、質感はモデリングデータに張り付けられるテクスチャや色によって表現される。本実験では、特徴ベクトルの直交性を保証するために、モデリングデータや使用

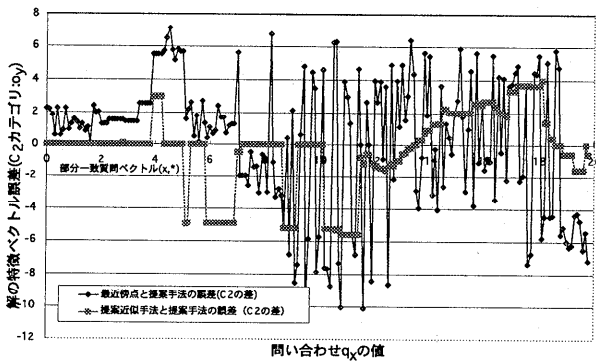


図 10 シミュレーション結果 (問い合わせの変更)
Fig. 10 Simulation Results(Change of query)

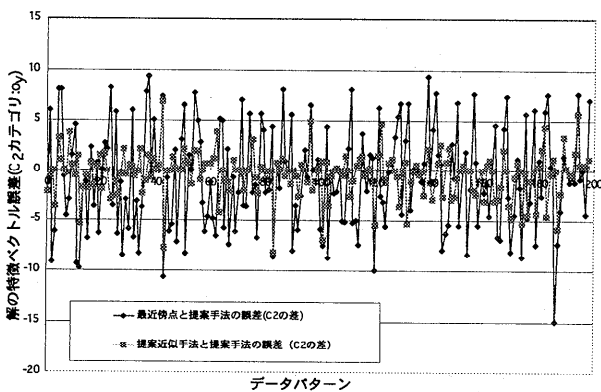


図 11 シミュレーション結果 (データの変更)
Fig. 11 Simulation Results(Change of data)

されているテクスチャ情報, 家具の特性を用いて形状に関する特徴, 色彩に関する特徴, 材質に関する特徴から独立した要素を抽出し, それらを用いて特徴を記述する 8 次元の特徴ベクトルを決定した. そして, 各要素オブジェクトへの特徴記述は, 特徴ベクトルの各要素に対して, その適合度を 0 から 8 までの整数値を用いて記述した. 決定した特徴ベクトルの各要素を以下に示す.

- 形状に関する特徴

- 直線的な部分が多い ↔ 曲線的な部分が多い

形状データの曲線と直線の量を尺度とする.

- 装飾的な構成要素が少ない ↔ 多い
- オブジェクトの基本構成部品以外の装飾的な部品の数を尺度とする.

- 重厚な作り ↔ 軽薄な作り
- オブジェクトの基本構成部品の太さや厚みを尺度とする.

- 色彩に関する特徴: 机, 椅子, その他の各カテゴ

リにおいて視覚的に重要である構成部品のテクスチャや色に着目し, HSV 表色系に基づいた Hue(色相), Saturation(彩度), Value(明度) 情報を特徴ベクトルの適合度の尺度とする.

- 色相: 暖色を多く含む ↔ 寒色を多く含む
- 椅子の座面, 机の天板等のように重要な構成部品の色に着目し, RGB 値 (0,255,0) の緑を中間値として上で, 青色方向を寒色, 赤色方向を暖色とした HSV 色相情報を尺度とする.
- 明度: 白っぽい ↔ 黒っぽい
- 各構成部品の最も多い色の明度を適合度の尺度とする.
- 彩度: 彩度の低い ↔ 彩度の高い
- 各構成部品の最も多い色の彩度を尺度とする.
- 色のコントラストが弱い ↔ 色のコントラストが強い
- 複数の色が含まれる場合にそれらの色相情報の差を尺度とする.

- 材質に関する特徴: 今回収集した家具のデータは大別すると木製家具のデータと一部木製あるいは木製部分のない家具データに分類可能であるため, 材質の特徴として木製部品の数に着目した.

- 木製部分が少ない ↔ 多い

6.2 編集オブジェクトの作成と登録

机, 椅子, その他のカテゴリに含まれる要素オブジェクトの中から, 各記述者が主観的に「ふさわしい」と感じる組み合わせを自由に作成した. 作成した組合せによって編集オブジェクトが作成され, その特徴は 24 次元の特徴ベクトルで表される. 本実験では, 図 12 のインターフェースを用いて, 特徴ベクトルの記述, 要素オブジェクトを組み合わせた編集オブジェクトの制作を行い, データベースに登録した. 編集オブジェクトは複数の人間が制作するため, 同じ構成を持つ編集オブジェクトの登録を許した上で, 100 個の編集オブジェクトをデータベースに登録した.

6.3 実験方法

本手法で提案した組合せ傾向を考慮した部分一致検索手法の有効性を検証するための実験を以下の様な手順で行った.

- データベースの中から机カテゴリ内の要素オブジェクト *desk*, 椅子カテゴリ内の要素オブジェクト *chair*, その他のアイテムカテゴリの *item* から構成される編集オブジェクト $O_{edit} = (desk, chair, item)$ を検索のための基本オブジェクトとして選択し, 椅子カテゴリの要素オブジェ

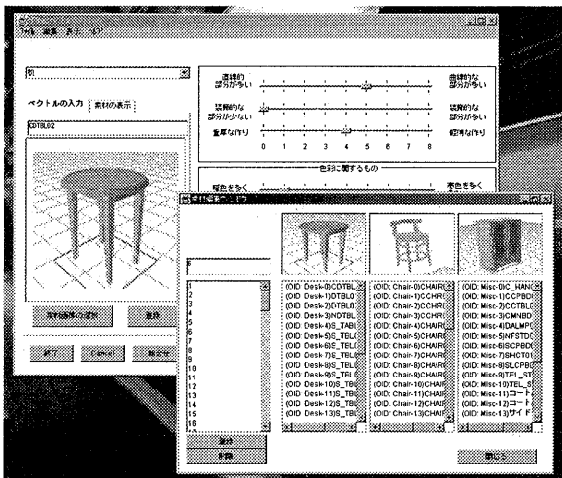


図 12 3D 素材検索システム

Fig. 12 3D asset retrieval system

クトを問い合わせる部分一致問い合わせ $q = (desk, *, item)$ を作成する。

- 基本オブジェクト o_{edit} を構成する要素オブジェクトの中から、椅子カテゴリの要素オブジェクトを「ふさわしくない」と考えられる要素オブジェクト $chair'$ に変更した特異な組合せとなる特異編集オブジェクト $o'_{edit} = (desk, chair', item)$ を作成し、データベースへ登録する。
- 部分一致問い合わせ q を行い、最近傍検索手法、提案手法の両手法における基本編集オブジェクトと特異編集オブジェクトの順位、最も問い合わせにふさわしい組合せであると評価された編集オブジェクトと特異な編集オブジェクトの類似度差について評価する。

以上の実験をデータベースに登録されてある 100 個の編集オブジェクト全てについて行った。また、類似度計量関数においては、前章のシミュレーション実験と同じ類似度計量式を使用した。

6.4 検索結果の評価

実験の一例として図 13 に最近傍検索の例を示す。この例では、特異な編集オブジェクトと、その特異な編集オブジェクトを作成する元になった基本編集オブジェクトの両方が同一の類似度を示し、解候補順位最上位となった。実験した全ての問い合わせにおいて同様に、最近傍検索手法では特異な編集オブジェクトが解候補最上位となり、最も問い合わせに類似度が高い解の一つであると判断された。

次に提案手法の実験例を図 14 に示す。この時、基本編集オブジェクトは、解候補順位最上位として選択

され、特異な編集オブジェクト解候補順位が 6 順位降下している。このように提案手法では、データベース内の編集オブジェクトの組合せ傾向を考慮することにより、適切な編集オブジェクトが選択できる。

このような実験をデータベースに登録されている 100 個の編集オブジェクトに対して行った。この時提案手法においては、解候補の最上位となった編集オブジェクトの類似度平均は 0.035 であった。また、最近傍検索手法において、必ず解候補の最上位となっていた特異な編集オブジェクトは、図 14 の実験結果のように提案手法において解候補順位が低下することが 100 件中 90 件認められた。この時の解候補順位降下度の平均は 5.9 であり、提案手法で得られた解候補最上位の編集オブジェクトと特異編集オブジェクトとの類似度差の平均値は、0.009 であった。また、提案手法における特異編集オブジェクトの類似度の平均値は 0.026 であった。以降、これらの解候補順位降下度、類似度差を評価対象として用いる。

提案手法においては、問い合わせに類似した組合せ傾向を重視した検索を行うことが目的であり、問い合わせに類似した編集オブジェクトの数が多ければ多いほど、その組合せ傾向が反映された重み付け平均ベクトルが作成される。そして、その結果として作成された重み付け平均ベクトルと問い合わせに対してふさわしい組合せを持つ編集オブジェクトの特徴ベクトルは、より一層類似度が大きくなると考えられる。従って相対的に、特異な組合せを持つ編集オブジェクトは、解候補順位が低下し、特徴ベクトルの類似度も低下することが予想される。つまり、問い合わせに対してある一定の類似度以上の編集オブジェクトの数と特異な編集オブジェクトの解候補順位降下度、特異な編集オブジェクトの類似度差には相関関係があると考えられる。

そこで、全ての検索結果から、検索者が主観的にふさわしい組合せを持つと考えられる編集オブジェクトの類似度の最低値を調査した。そして、その平均値 0.025 を閾値として設定し、それ以上の類似度を持つ編集オブジェクトの数と特異な編集オブジェクトの類似度順位降下度と類似度差の相関係数を統計的処理により算出した。これは、問い合わせに類似した編集オブジェクトの数が、提案手法ではどのように影響するかを検証することが目的である。

類似度が閾値以上の編集オブジェクトの数と特異な編集オブジェクトの順位降下度、類似度差に対して算

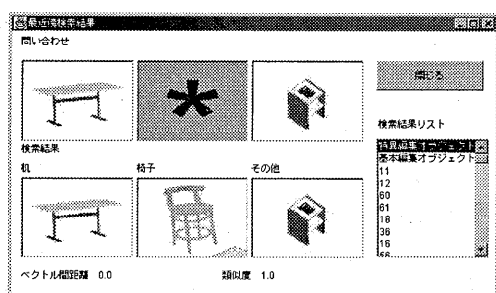


図 13 最近傍検索結果

Fig. 13 A Result of Nearest neighbor retrieval Method

出された相関係数はそれぞれ、0.37、0.45であり、問い合わせベクトルに類似した編集オブジェクトの数と特異な編集オブジェクトの解候補順位降下度にて正の相関があることがわかる。また、類似度差にも正の相関関係を持つことがわかる。つまり、問い合わせに類似したオブジェクトの数が多ければ多いほど、特異な編集オブジェクトを解候補から類似度を低下させて排除していく傾向が見られ、提案手法が目的とする過去の組合せ傾向を考慮した部分一致検索手法が有効に機能していることが確認された。

しかし、相関係数はそれほど大きいものではなく弱い相関関係しか現れていない。問い合わせに類似した編集オブジェクトが多いにも関わらず、特異な編集オブジェクトの解候補順位が上位になっているものを調べてみると、閾値として設定した類似度以上ではあるが、全体的に類似度が低く閾値付近に編集オブジェクトが多く分布しているものであった。従って、相関関係が弱い原因としては類似度計量関数で算出された値が小さく有効な重み付け平均ベクトルが算出されていなかったことが挙げられる。また、最近傍手法と提案手法、どちらの手法においても特異な編集オブジェクトが解候補の最上位となった検索結果が100回の実験中10件見られた。これらのケースにおいても、問い合わせに対して低い類似度の編集オブジェクトの分布がなく、類似度計量関数が算出した類似度の値が小さかったために、有効な重み付け平均ベクトルが算出されていなかった事が原因として考えられる。

7. おわりに

複数の要素オブジェクトを組み合わせで作成される編集オブジェクトに対するベクトル空間モデルを用いた部分一致検索の手法について述べた。この手法は、問い合わせベクトルに対して類似度の高い編集オブジェクトに含まれる要素オブジェクトの組合せ傾向を考慮することで、部分的な問い合わせにおけるふさわ

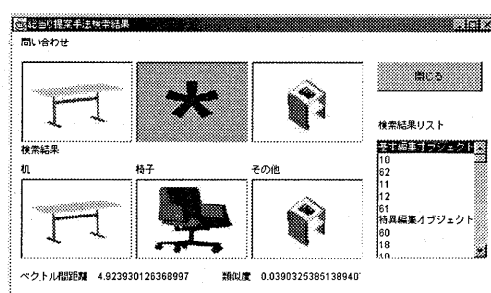


図 14 提案検索結果

Fig. 14 A Result of Proposed Retrieval Method

しい組合せを検索することが可能である。また、計算量を削減した近似手法についても提案し、シミュレーションを用いて提案手法の有効性を確認した。更に、3D素材を対象とした検索システムを開発し、マルチメディアデータの部分一致検索に適用し、提案手法の有効性を確認した。

この手法は、文書データベースなどの一般的なベクトル空間モデルを用いた検索手法にも適用可能である。文書データベースや印象語を用いた検索を行う場合には、特徴ベクトルの直交性を保証することが必要になるが、既に提案されている手法を用いることで独立した要素からなる特徴ベクトルが作成できれば、提案手法は有効に機能する。

しかし、オブジェクトの類似度を計算する任意の評価式において、最適な評価式を設定することは、容易ではない。第6章で行った実験においても、多くの類似した編集オブジェクトが存在する場合には提案手法の有効性が確認できたが、問い合わせに対して類似度が低い編集オブジェクトしか存在しない場合には、適切な重み付け平均ベクトルを算出することが困難であった。

従って、最大限の効果を得るための評価式の発見については、今後の課題である。今回は、発見的に類似度計量関数を設定したが様々なデータ分布や更に高次元な特徴ベクトル空間において、最適な結果が得られる類似度計量関数の設計指標を構築していく必要がある。

今後、高次元のベクトル空間モデルでの評価、文書データベースなどへ適用し、類似度計量関数の設計指標の確立を行いながら提案手法の改良を行う予定である。

謝辞 本研究は、通信・放送機構神戸リサーチセンターにおける次世代デジタル映像通信に関する研究の一貫として行われた。この研究を行うにあたって実験

素材の準備, 実験補助等にご協力戴いた神戸大学情報知能工学科の松本 好市氏に謝意を表す. なお, 著者の一部(田中)は本研究において, 一部, 文部省重点領域研究(課題番号 8244103)及び, 日本学術振興会未来開拓学術研究推進事業における研究プロジェクト「マルチメディア・コンテンツの高次処理の研究」によっている. ここに記して謝意を表す.

参 考 文 献

- 1) K.Hatano, T.Kamei and K.Tanaka: Clustering and Authoring of Video Shots using Hybrid-type Self-Organizing Maps, Proceedings of International Symposium on Digital Media Information Base (DMIB'97), pp.150-158, November (1997)
- 2) Y.Kidawara, F.Sawa, M.Kawauchi and K.Tanaka: Authoring and Retrieval of Digital Assets for Virtual Studio System, Proceedings of International Symposium on Digital Media Information Base (DMIB'97), pp.11-19, November (1997)
- 3) C.Faloutsos: Searching Multimedia Databases by Contents, Proceedings of International Symposium on Digital Media Information Base (DMIB), pp.204-205, November (1997)
- 4) W.Chang, G.Sheikholeslami and A.Zhang: Efficient Resource Selection in Distributed Visual Information Systems, Proceedings of ACM Multimedia '97, pp.203-213, (1997)
- 5) C.Faloutsos, K.Lin: FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets, Proceedings of ADM SIGMOD '95, pp.163-174, (1995)
- 6) D.Schaffer, Z.Zuo, S.Greenberg, L.Bartram and M.Roseman, ACM Transactions on Computer-Human Interaction Vol.3, No.2, pp.162-188 (1996)
- 7) Christos Faloutsos: Searching Multimedia Databases By Contents, Kluwer Academic Publishers, ISBN 0-7923-9777-0
- 8) Y. Kiyoki, T. Kitagawa and T. Hayama, "A Metadatabase System for Semantic Image Search by a Mathematical Model of Meaning," Multimedia Data Management - using metadata to integrate and apply digital media-, McGrawHill(book), A. Sheth and W. Klas(editors), Chapter 7, March, 1998.

(平成 10 年 9 月 20 日受付)

(平成 11 年 1 月 2 日採録)

(担当編集委員 加藤 俊一)

木俣 豊(正会員)



1988 年 神戸大・工・計測工学卒, 1990 年 同大学大学院工学研究科修士修了. 同年株式会社神戸製鋼所入社. 製鉄所生産管理データベース, 自動車用 ABS 自動計測装置, 自動車用 ABS リアルタイムシミュレータの研究開発に従事. 現在, 通信・放送機構神戸リサーチセンター出向し, 次世代映像デジタル通信技術の研究開発プロジェクトに従事, 神戸大学大学院自然科学研究科(情報メディア科学専攻)博士後期課程在学中. 第 54 回情報処理全国大会優秀賞受賞, 情報処理学会, システム制御情報学会等各会員

澤 扶美(学生会員)



1998 年 神戸大・工・情報知能工学卒, 現在, 同大学大学院自然科学研究科修士課程在学中. 主にマルチメディア情報検索に関する研究に従事.

田中 克己(正会員)



1974 年 京大・工・情報工学卒, 1976 年 同大学大学院修士修了. 1979 年神戸大学教養部助手, 1986 年 同大学工学部助教授. 1994 年 同大学工学部教授(情報知能工学専攻). 1995 年 同大学大学院自然科学研究科(現在, 情報メディア科学専攻)専任教授, 現在に至る. 工博. 主にデータベースの研究に従事. 現在, 情報処理学会データベースシステム研究会主査. 96 年度より通信・放送機構「次世代デジタル映像通信の研究開発」の研究統括責任者, 文部省科研費重点領域研究「分散発展型データベースシステム技術の研究」の研究代表者, 神戸マルチメディアインターネット協議会会長, 情報処理学会, 人工知能学会, IEEE Computer Society, ACM 等各会員.