D-002

類似検索機能を有する灰色理論型データベースの一提案
Grey Database System with Similarity Retrieval

山口大輔 † 小林俊裕 † 水谷晃三 † 赤羽根隆広 † 張雪元 § 永井正武 †‡
Daisuke Yamaguchi Toshihiro Kobayashi Kozo Mizutani
Takahiro Akabane Xue-Yuan Zhang Masatake Nagai

1. はじめに

検索キーと一致しているデータだけを出力する検索方法を直接検索と定義した場合，完全に一致していなくても，問い合わせ内容と概ね一致していれば出力する検索方法を類似検索という。イメージを基に検索する方法は主に後者を利用する。

マルチメディアに代表される様々なコンテンツの検索は，個人の主観性に大きく依存している。そのため，個人の主観的イメージに合ったオブジェクトを類似検索するためのコンテンツ構造化手法ならびに検索手法が多数提案されている [1]。

本研究は利用者のイメージにより検索可能で，かつ広分野に適用可能なデータベースの開発を目的とする。本稿は灰色理論による数理的類似検索アルゴリズムを持つデータベースの構成技法を提案する。

2. 灰色理論

灰色理論 [2, 3] は，ある主題に対する情報の部分既知状態を数理的に完全既知状態にすることを目的とする基礎数理理論である。本稿では灰色理論に含まれる分析手法である灰色分析を使用する。灰色分析とは，複数の数列間におけるデータの類似度を灰色関連度 $\Gamma \in [0,1]$ の実数値として出力する分析法である。いま，X を灰色関連因子集合とすれば，その原始数列 x_i は

$$x_i = \{x_i(1), x_i(2), \ldots, x_i(k)\} \in X \qquad (1)$$

とする。したがって，数列とは複数の観測値からなる集合と定義する。式 (1) は観測値を持ち，類似度を測りたい数列である。この数列 $x_i(i = 1, 2, \cdots, n)$ を比較数列と定義する。

そして，類似度を測る基準となる数列，基準数列 x_0 を 1 組用意し，以下に定義する。

$$x_0 = \{x_0(1), x_0(2), \ldots, x_0(k)\} \qquad (2)$$

最初に，基準数列の要素 $x_0(j)$ と各比較数列の要素 $x_i(j)$ との差の絶対値 $\Delta_{0i}(j)$ を算出する。

$$\Delta_{0i}(j) = |x_0(j) - x_i(j)| \quad (j = 1, 2, \ldots, k) \qquad (3)$$

次に，灰色関連係数 $\gamma_{0i}(j)$ を次式から計算する。

$$\gamma_{0i}(j) = \frac{\min\limits_{\forall i}\min\limits_{\forall j}\{\Delta_{0i}(j)\} + \zeta \max\limits_{\forall i}\max\limits_{\forall j}\{\Delta_{0i}(j)\}}{\Delta_{0i}(j) + \zeta \max\limits_{\forall i}\max\limits_{\forall j}\{\Delta_{0i}(j)\}} \qquad (4)$$

ただし，灰色分析の調節用係数として $\zeta = 0.5$ とする。

次式の処理により，分析結果である灰色関連度 Γ_{0i} を比較数列ごとに得ることができる。

$$\Gamma_{0i} = \frac{1}{k}\sum_{j=1}^{k}\gamma_{0i}(j) \quad (i = 1, 2, \ldots, n) \qquad (5)$$

$\Gamma_{0i} \to 1$ であるほど数列 x_0,x_i 間のデータは類似していることを表す。上述の灰色分析は，本稿で提案するデータベースの数理的拠点となる。

3. 灰色理論型データベース概要

本稿で提案するデータベースは，坂井ら [4] の報告による感性データベース概念の流れをくむものとする。坂井らは因子分析やクラスター分析を用いた多変量解析によるデータベースを提案している。本提案データベース構成技法は，灰色理論による類似検索機能，データ追加機能，テーブル再構成機能といった，効率的な検索を行うための機能を有することが特徴である。以下に本提案データベースの主な機能について述べる。

3.1 データベース構成法

図 1 に本稿で提案するデータベースの構成方法を示す。数理データベースであることを前提として，取り扱うデータは任意の実数とする。本データベース構成法は，はじめにデータを収集し，灰色分析を応用したクラスタリング [3, 5] を行う。

データベース化する n 件のデータをクラスタリング対象数列 $d_i(i = 1, 2, \ldots, n)$ とする。全 d_i を比較数列，各 d_i を一度ずつ基準数列に設定し，n 回灰色分析を行う。その結果，次式に示す灰色関連マトリクス $R_{n\times n}$ が得られる。ただし，d_i を基準数列としたときの比較数列 d_j

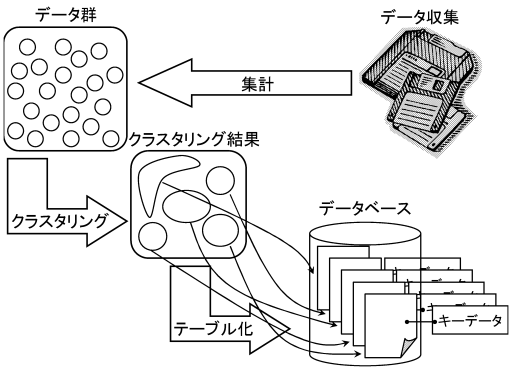

図 1 灰色理論型データベース構成法

†帝京大学大学院理工学研究科，Graduate School of Sciense and Engineering, Teikyo University
‡帝京大学理工学部，School of Sciense and Engineering, Teikyo University
§北京科技大学，University of Sciense & Technology Beijing

との灰色関連度を r_{ij} とする。

$$R_{n\times n} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & r_{ij} & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix} \tag{6}$$

2 つの灰色関連度 r_{ij} と r_{ji} から，2 件のデータ d_i, d_j 間の類似度である灰色相関係数 $l(d_i, d_j)$ を次式にて算出する。

$$l(d_i, d_j) = \frac{r_{ij} + r_{ji}}{2} \tag{7}$$

ただし，

$$\begin{cases} i < j \\ i = 1, 2, \ldots, n-1 \\ j = 2, 3, \ldots, n \end{cases}$$

とする。

灰色相関係数を算出後，$l(d_i, d_j)$ を降順整列し，$\max_{\forall l} l(d_i, d_j)$ となるデータ d_i, d_j から順にデータをクラスタリングしていく。

新たに構成されるクラスターを t とすると，t はクラスタリング対象データ d および既存のクラスター p, q の集合と捉えることができる。クラスター t は，

$$t = \begin{cases} p & (d_i \in p \wedge d_j \in p) \\ p \cup q & (d_i \in p \wedge d_j \in q) \\ d_i \cup p & (d_j \in p) \\ d_i \cup d_j & \text{otherwise} \end{cases} \tag{8}$$

により構成される。これらのクラスタリング処理を繰り返し，m 個のクラスターが構成されたところで処理を終了する。

上記のクラスタリングアルゴリズムにより，類似傾向にあるデータのグループを数理的に獲得する。1 グループごとに 1 テーブルとしてデータベースに格納する。類似傾向のデータを 1 テーブルとすることで検索効率向上を図る。

データベース構成後，各テーブルを識別するための代表データを獲得する。この代表データのことをキーデータと定義する。キーデータは各テーブル内のデータの平均値などを設定する。

3.2 類似検索機能

図 2 に示す検索法は本稿で提案する類似検索機能である。入力として，利用者の主観的なイメージを数値データとして獲得し，検索キーとする。データベース内部では，検索キーを基準数列 x_0，テーブル $T_j (j = 1, 2, \ldots, m)$ におけるキーデータを比較数列 x_j として 2. で述べた灰色分析を行う。灰色分析の結果，

$$\max_{\forall j} \{\Gamma_{0j}\} = T_j \equiv T_{select} \tag{9}$$

となるテーブル T_{select} を選択する。次にデータ検索を行う。利用者が入力した検索キーを基準数列 x_0，T_{select} 内の各データ $d_i (i = 1, 2, \ldots, n)$ を比較数列 x_i として灰色分析する。灰色分析の結果，出力するデータの集合 O は

$$O = \{d_i | d_i \in T_{select}, \Gamma_{0i} \geq w\} \tag{10}$$

である。ただし，w は任意の閾値で，$0 \leq w \leq 1$ とする。

灰色関連度 Γ_{0i} の上位に位置するデータ d_i を利用者の要求に最も近いデータとして出力する。したがって，出力するデータの集合 O は，d_i を Γ_{0i} について降順整列した，

$$O = \{d_i, \ldots, d_j\} \quad \text{if } (\Gamma_{0i} \succ \cdots \succ \Gamma_{0j}) \tag{11}$$

となる。以上のアルゴリズムにより類似検索機能を実現する。

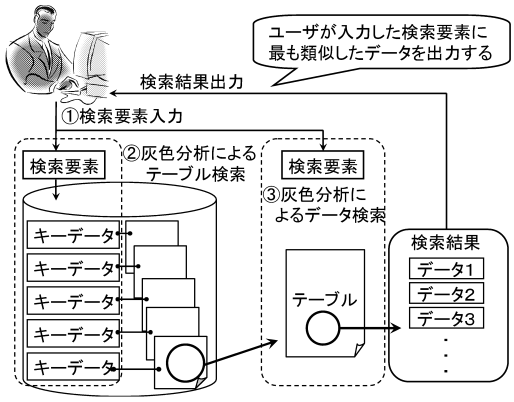

図 2 灰色理論型データベースによるデータ検索

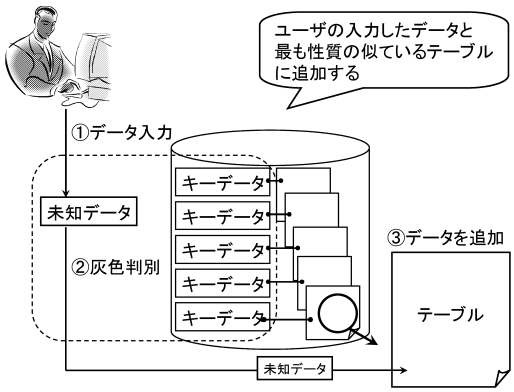

図 3 灰色理論型データベースによるデータ追加

3.3 データ追加機能

図 3 にデータ追加機能を示す。本機能は，追加するデータが複数あっても一括して追加処理を行うことが可能である。利用者は追加するデータを数値入力する。入力されたデータはデータベースにとって未知データである。そこで，灰色分析を応用した判別モデル [3, 6] を用いて入力データを追加する。

追加データ n 件を比較数列 $c_i (i = 1, 2, \ldots, n)$，テーブル $T_j (j = 1, 2, \ldots, m)$ におけるキーデータを判別基準数列 s_j とする。s_j を基準数列としたときの追加データ c_i

との灰色関連度を a_{ji} とすると，全 s_j を用いて m 回灰色分析した結果は，次式の行列 A で表すことができる。

$$A = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & a_{ji} & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{pmatrix} \quad (12)$$

各追加データ c_i は，次式のルールによっていずれかのテーブル T_j に追加する。

$$T_j(c_i) = \max_{\forall j}\{a_{ji}\} \quad (i = 1, 2, \ldots, n) \quad (13)$$

上記に示す灰色判別モデルのアルゴリズムにより，各データは m 個あるテーブルのうちいずれか 1 つに追加する。追加データはデータの性質が最も似ているテーブルに追加されることになる。

3.4 テーブル再構成機能

データを追加していくうちにデータベースは肥大化する。そこで，定期的にテーブルの再構成を行い，検索精度の向上を図ることが容易に行える。テーブル再構成機能は，全テーブルを開放し，全件のデータに対して 3.1 に示すテーブル構成処理を再度適用する。

4. 評価実験

本提案データベースでは，灰色分析を応用したクラスター分析および判別モデルを用いている。伝統的な統計手法を用いても本稿で提案している機能を持つデータベースの実現は可能である。そこで，本提案方法が統計的手法で実現した場合よりも有益かどうか評価を行う。

4.1 実験内容

灰色理論によるデータベース及び統計手法によるデータベースの検索精度および処理効率を実験により比較する。本稿では以下の 4 つの実験を行う。

1. 検索精度の測定
   前提条件　検索パラメータ数 k の検索で得られた第一候補データとの相対誤差を算出する。ただし，$k = \{4, 5, 6, 7\}$ とする。

2. データ検索所要時間の測定
   前提条件　1 テーブル z 件のデータから，利用者の問い合わせに対する 10 件のデータを出力する。ただし，$z = \{50, 100, 200, 300, 500\}$ とする。

3. データベース構成所要時間の測定
   前提条件　x 件のデータをクラスタリングし，10 テーブルを構成する。ただし，$x = \{500, 1000, 2000, 3000, 5000\}$ とする。

4. データ一括追加所要時間の測定
   前提条件　y 件のデータを 10 テーブルのいずれかに追加する。ただし，$y = \{100, 300, 500, 700, 1000\}$ とする。一括処理に要した時間を測定する。

表 1 比較対象データベースのアルゴリズム

データベース構成法	ウォード法によるクラスタリングを行う。クラスタリング後，各テーブルのキーデータを設定する。
データ追加方法	追加データとキーデータとのユークリッド距離を計算し，最も距離の小さいテーブルに追加する。
データ検索方法	検索データとキーデータとのユークリッド距離を計算し，最も距離の小さいテーブルを 1 つ選択する。選択されたテーブル内部のデータと検索データとのユークリッド距離を計算し，距離の近いデータの上位を出力する。

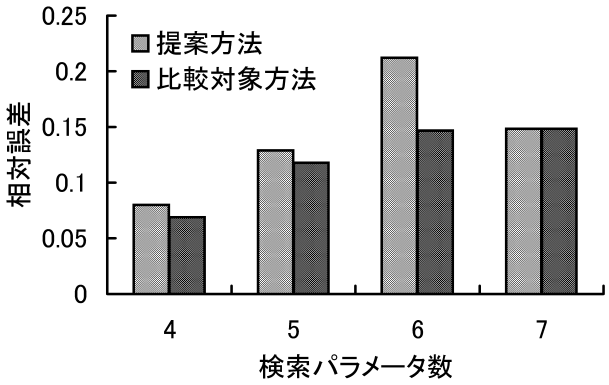

図 4 検索精度測定結果

4.2 実験方法

本実験は，統計手法によるクラスター分析を用いたデータベースを比較対象とする。比較対象データベースのアルゴリズムの概略を表 1 に示す。取り扱うデータは，任意の [1,100] の値とする。実験 1 は相対誤差を比較して検索精度を検討する。実験 2 は 3.2 による提案方法と表 1 に示すデータ検索方法の両者の所要時間を比較する。実験 3 は 3.1 にて示した提案方法と表 1 に示すデータベース構成法の両者における所要時間を比較する。実験 4 は 3.3 にて示した提案方法と表 1 に示すデータ追加方法の両者の所要時間を比較する。

4.3 実験結果

図 4 は実験 1 の測定結果，図 5 は実験 2 の測定結果である。図 6 は実験 3 の測定結果，図 7 は実験 4 の測定結果である。検索精度は比較対象手法の方が良く，データ追加時間は両方法とも同程度の結果である。データベース構成所要時間は，提案方法の方が約 29%短い結果となっている。一方，データ検索時間では，比較対象方法が提案方法よりも約 42%短い結果となっている。

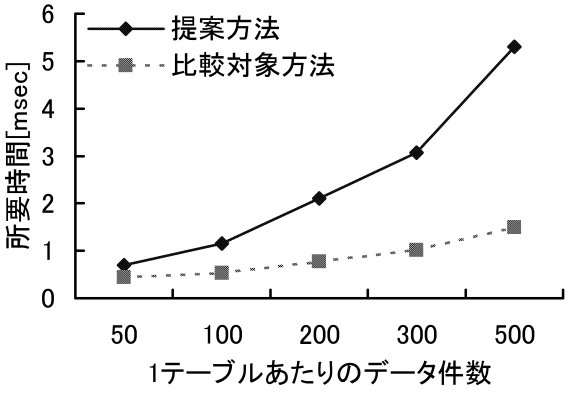

図 5 データ検索所要時間測定結果

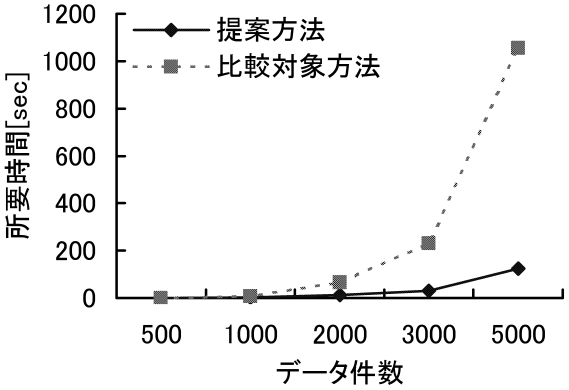

図 6 データベース構成所要時間測定結果

5. 考察

5.1 提案手法の評価

実験により，検索精度は提案手法の誤差が大きいことがわかった。これは，灰色分析のアルゴリズムにおいて重みが均等でないことを表している。1 つでも $\gamma_{0i}(j)$ の高いパラメータが存在すると，他の $\gamma_{0i}(j)$ が低い値でも式 (5) による平均値計算により全体的な類似度が底上げされてしまうからである。改善案として，式 (5) に重みを均等にする方法 [3] を導入することが考えられる。

検索所要時間は，提案方法が比較対象方法よりも約 42%効率が悪い結果となっている。これは，灰色分析よりもユークリッド距離計算の方が早いことに起因する。

5.2 期待できる適用分野

本稿で提案する類似検索はあいまい検索や感性検索に相当する。期待される適用分野における用途例を以下に挙げる。

(1) マルチメディアデータベース

画像・映像・音楽・配色・服装のコーディネートなど，利用者のイメージでコンテンツを検索するシステムの実現が期待される。本稿で提案するデータベースが感性検索に応用できる。

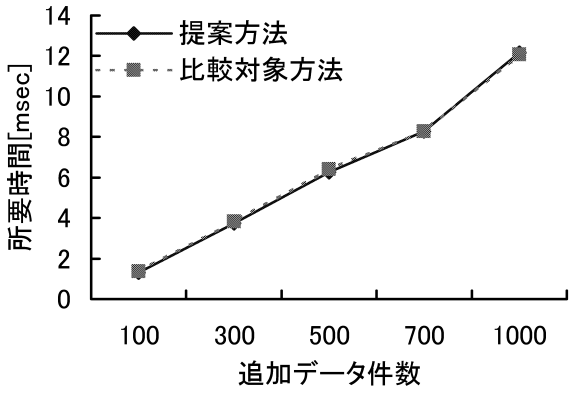

図 7 データ一括追加所要時間測定結果

(2) 商品データベース

様々な商品における仕様などを数値データとしてデータベース化する。仕様の類似している商品を各テーブルに配置することで商品の特徴を整理したデータベースを構築することができる。インターネットショッピングなどにおける商品のあいまい検索に本データベースは適用できる。

(3) 個人情報データベース

個人の趣味や嗜好を数値データとして獲得し，類似傾向にある個人同士をグループ化する。メーリングリストやユーザグループを構成し，活用・管理をするのに本データベースは利用可能である。

6. おわりに

本稿は灰色理論による数理的な類似検索機能やテーブル構成機能を持つデータベースを提案した。統計的手法で本提案データベースと同等の機能を実現し，検索精度や処理効率を実験にて比較および検討を行った。

今後は本提案データベースを具体的なマルチメディアコンテンツに適用し，利用者の満足度について検証していく予定である。

参 考 文 献

[1] 西尾章治郎, 田中克己, 上原邦昭, 有木康雄, 加藤俊一, 河野浩之, 情報の構造化と検索, 岩波書店, 2000.

[2] 鄧 聚龍 (著), 趙 君明, 北岡正敏 (訳), 灰色理論による予測と意思決定, 日本理工出版会, Dec. 1999.

[3] 永井正武, 山口大輔, 灰色理論と工学応用方法入門, 共立出版, 2004.（出版予定）

[4] 坂井伸明, 大塚真吾, 宮崎収兄,“多変量解析を用いた感性データベース,” 信学技報, DE2001-99, pp.159–166, July 2001.

[5] 山口大輔, 小林俊裕, 水谷晃三, 永井正武,“灰色分析を適用した階層的クラスター分析法の提案,” 情処研報, vol.2004, no.10, AL-93-11, pp.75–82, Jan. 2004.

[6] 山口大輔, 小林俊裕, 水谷晃三, 永井正武,“灰色分析を適用した判別モデルの提案,” 信学技報, vol.103, no.521, HIP2003-82, pp.31–36, Dec. 2003.