データベースシステム　124－4
情　報　学　基　礎　　62－4
（２００１．５．２１）

多様な属性表現に対応した類似検索の一手法

村本　達也　　池田　哲夫
ＮＴＴサイバースペース研究所
muramoto@isl.ntt.co.jp, ikeda@dq.isl.ntt.co.jp

類似検索手法の多くは，検索キーオブジェクトと検索対象オブジェクトを n 次元空間上の点へそれぞれマッピングしてこの点間距離に基づき類似度を算出している．しかし時空間情報など「域」で表現される属性を含んだデータの検索において，「域」で表現される情報を「点」に縮退させると属性が表現する性質を部分的にしか検索結果に反映できない．本稿では，「点」で表現される属性や「域」で表現される属性の双方の属性表現を含むデータベースに対して，各属性の特徴量を「域」で管理し，各属性が表現する領域の積集合領域と和集合領域の比を用いて類似度の計算を行う手法を提案する．また，従来の「点」へ縮退させて点間距離に基づき類似度を計算する手法との比較実験を行い，提案手法の有効性を示す．

A Similarity Search Method Adapted to Flexible Data Representation

Tatsuya MURMAOTO and Tetsuo IKEDA
NTT Cyber Space Laboratories
muramoto@isl.ntt.co.jp , ikeda@dq.isl.ntt.co.jp

In this paper, we propose a novel similarity search method adapted to flexible data representation. Most of traditional similarity search methods calculate the similarity based on distance in *n*-dimension space. The characteristic of the proposed method is the treatment of the attributes whose value is not a point but an area or an interval. The similarity is calculated using ratios of the intersection of attribute value(area) of query object and that of target object to the union of them. Through the comparison of our method to the Distance-based retrieval method, the effectiveness of our method was proved.

1. はじめに

計算機の高性能化と低価格化に伴い，文書，静止画，動画，音声等の多種多様な情報が生成され，流通しつつある．これらの多種多様な情報の中から必要な情報を容易に入手可能とするために様々な検索手法が提案されている[1][2]．このような検索手法においては，類似度に基づく検索の仕組みが中心となっている．

類似度に基づく検索手法の多くは，検索キーオブジェクトと検索対象オブジェクトを複数の特徴量で表現して多次元空間上の点（ベクトル）にマッピングし，その点間の距離をもって類似度を計算して，検索キーオブジェクトがマッピングされた点に最も距離が近い点で表現されるオブジェクト群を探索する最近傍点探索法に基づいている．また，距離の定義方法をユークリッド距離ではなく楕円体距離とすることで，利用者の直感により適応した検索を実現する手法も提案されている[3]．この点間距離に基づく方法は，文書検索をはじめ，パターン認識，統計解析や，画像，音楽などのマルチメディアデータベース検索を中心に広く利用されている．

この最近傍点探索法ではオブジェクトが持つ各属性の特徴量を「点」で表現することを特徴としている．しかし，時空間情報など「域」で表現される属性を含んだデータの検索において，「域」で表現される情報を「点」に縮退させてしまっては属性が表現する性質を部分的にしか反映することが出来なく，妥当な検索結果を得られない可能性があるものと筆者らは考える．「域」で表現される特徴量は図1中の (a) の様な単体の区間で表現できる場合だけではなく，(b) の様に特徴量が不連続な区間群で表現される場合や，(c) の様に複数の属性が互いに関連していて複数の領域の組み合わせで表現される場合も存在する．特に (b) (c) の場合に「点」へ縮退させると，多くの情報が欠落して検索精度を落とすと考えられる．以下本稿では，(b) の様に特徴量が不連続な区間群で表現される場合を「分割表現」，(c) の様に複数の領域の組み合わせで表現される場合を「従属表現」と呼ぶこととする．また，1 次元の区間，2 次元以上の領域を本稿では区別せず「領域」と呼ぶこととする．

本稿では，上記で挙げたような多様な属性表現をふくむデータベースに対して，「域」で表現された情報を「点」に縮退させることなく，精度の良い類似度計算を行うことの出来る手法を提案する．

以下 2 章では従来の類似検索手法の問題点を挙げ，3 章では多様な属性表現を含むデータベースに対する類似度計算手法を提案する．4章では提案手法を筆者らが研究を進めている地図連動型情報提供サービスへ適用したプロトタイプを説明し，5章では本手法の評価を行う．

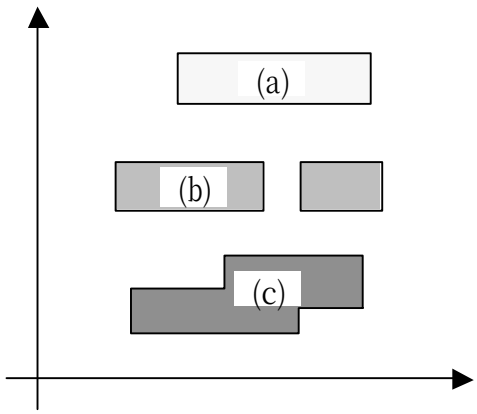

図 1：多様な属性表現の例

2. 従来手法とその問題点

従来方法では，特徴量がある一定の幅を持つ性質や分布で表現される性質を持つオブジェクトを対象とする類似検索を行う場合，その様な特徴量を重心などの代表値に縮退させるなどして各特徴量を点へマッピ

ングし，点間距離に基づく類似度計算を行っている．

しかしこの代表値に縮退させる方法では，特徴量の性質の一部を失ってしまい，妥当な検索結果が得られない可能性がある．例えば，午前 9 時から午後 5 時という様な「域」で表現される特徴量を「点」に縮退させると，領域の幅の情報が欠落して検索精度を落とす可能性がある．更に図1に示した様な，特徴量が分割表現や従属表現された場合は情報の欠落がより多くなる．分割表現の例として，営業時間を午前 10 時から午後 1 時と午後 4 時から午後 8 時というような領域群で表現する例が考えられる．この特徴量の性質を「点」に縮退させると，領域群を構成する各領域の数・幅などの情報は失われ，類似度計算の精度を悪化させると考えられる．また，従属表現の例として，営業時間が午前 9 時から午後 1 時までは価格が 1000 円から 3000 円，午後 1 時から午後 9 時までは 2000 円から 5000 円という例が考えられる．この複数属性の特徴量の性質を「点」に縮退させると属性間の関連情報を失うので類似度計算の精度が悪化すると考えられる．

3. 提案手法

前節で述べた従来方法の問題を踏まえ，本研究では各属性の特徴量を「点」へ縮退させることなく「域」で管理し，類似検索を行う手法を提案する．本節では先ず「域」データの管理方法について説明し，続いて提案する類似度計算手法について述べる．

3.1. 特徴量の管理方法

類似度計算式の単純化を考慮して各属性を一様な方法で管理するのが望ましいと筆者らは考える．そこで各属性毎の特徴量を領域の集合として管理することとする．

各属性毎の特徴量を領域の集合として管理することとする場合，以下に挙げる2つの場合の管理方法を検討する必要がある．

A. 従属表現される複数の属性の特徴量を管理する場合．

B. 「点」で表現される特徴量を「域」で管理する場合．

A. の場合，複数の関連した属性群は独立しておらず図1中(c)の様に複数の互いに異なる域で表現される領域の集合となるので，各属性毎に特徴量を管理しては，属性間の関連性を保持できない．そのため，従属表現を含む属性群に対しては，関連する k 個の属性群の特徴量を k 次元の領域の集合として管理する．

B. の場合に対しては，他の「域」で表現される属性との整合性を保つために，緯度・経度等，数直線上にマッピングできる性質の属性に対しては値に $\pm const$ の幅を与えて，ジャンルや性別などそのまま数直線上にマッピングしても意味の無い属性に対しては特徴量の性質を表現できる構造にマッピングを行う．例えば，ジャンルの様に多段階の階層構造で表現できるものに対しては，木構造のグラフの節点に各ジャンルをマッピングし，グラフの節点間の距離を定義する．この方法によって任意のジャンルを原点とした距離を数直線上に表現することが出来，$const$ の幅を与えて「域」の属性として特徴量を管理することが出来る．

この方法より各属性は一様に領域の集合で管理される．

3.2. 領域間の類似度計算手法の基本的考え方

領域間の類似度計算手法を導くに当たっては，図2に示すように領域間の重複部分の大きさと非重複部分の大きさに着目することとした．人間の類似度の直感に合致するのは，以下の要求条件を満たす式であると考える．

- 領域の重複部分が大きいほど類似度が大きくなる式．
- 領域の非重複部分が大きいほど類似度が小さくなる式．

この要求条件を踏まえて，検索キーオブジェクト K と検索対象オブジェクト X の類似度は，下記の(1)式の様に各属性毎で検索キーオブジェクトと検索対象オ

ブジェクトの領域群の積集合領域と和集合領域の比を求めこれを各属性毎の類似度とし，各属性毎の重み α_i を用いて α_i 乗根を求め，その積を取ることとする．

$$Sim(K,X)=\prod \sqrt[\alpha_i]{\frac{V(k_i)\cap V(x_i)}{V(k_i)\cup V(x_i)}} \qquad (1)$$

なお，従属表現を含む属性群に対しては，各属性間の関連を保持するために図 2 に示すように，多次元空間における $V(k_{n1},\cdots,k_{nm})$ ，$V(x_{n1},\cdots,x_{nm})$ を求め，この領域同士の積集合領域と和集合領域の比を属性群の類似度として用いることとする．

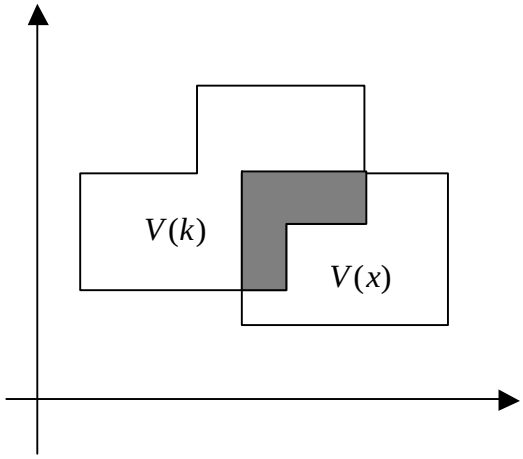

図 2：領域同士の類似計算のイメージ

3.3. 積集合が空の場合の対処

上記類似度計算式では各属性で積集合領域が空となると，検索キーの領域と検索対象の領域がどのぐらい離れていようと当該属性の類似度が0となる問題がある．この問題を回避する為の対処を本節で説明する．新たな評価式を導入し，前述の(1)式と組み合わせることによって回避することとする．組み合わせ方については次節で述べる．

評価式の要求条件を以下に挙げる．1点目は類似度計算という観点から必須の条件であり，2点目は(1)式と組み合わせることを考慮した場合に満たすと望ましい条件である．

領域同士が離れるに従い値が減少する式．

領域の積集合が空の時のみ値を変化させる式．

これらを踏まえて以下の評価式を導入する．

$$g(k_i,x_i)=\frac{1}{\max(1,D(V(k_i),V(x_i)))} \qquad (2)$$

ここで D は，領域間の距離関数であり，積集合領域が空の時のみ評価式 g の値へ影響させるために，領域同士の最近接点間距離を採用する．この距離関数は積集合領域が非空の状態ならば距離関数値は 0 を取るので空の状態のときのみ値を変化させることが可能になる．また，評価式の取り得る値を正規化するために，前述の距離関数と 1 との max を取る．

3.4. 類似度計算式

3.3 節で述べた評価式を 3.2 節の類似度計算式に組み入れた式を類似度計算式とする．組み入れる上で，類似度計算値を人間の直感と合致させるために，総ての属性で積集合領域が空になる場合の類似度が，積集合領域が非空となる場合の類似度を上回らないことが望ましい．

ここで，便宜上3.2節で述べた類似度計算式の積乗記号内部の部分を $f(k_i,x_i)$ とする．上記の要件を踏まえて，以下の様に属性ごとで $f(k_i,x_i)$ と $g(k_i,x_i)$ の線形和を取りその和を取る類似度計算式を提案する．

$$Sim(k,x)=\prod \sqrt[\alpha_i]{(\beta f(k_i,x_i)+(1-\beta)g(k_i,x_i))} \qquad (3)$$

$f(k_i,x_i)$ と $g(k_i,x_i)$ は，図3の様に関数値が変化するので，積集合領域が非空の場合には $f(k_i,x_i)$ のみ類似度計算に作用し，空の場合には $g(k_i,x_i)$ のみ作用する．よって，式(3)の様に $f(k_i,x_i)$ と $g(k_i,x_i)$ とを組み合わせれば，総ての属性で，積集合領域が空になる場合の類似度が非空となる場合の類似度を上回らない．

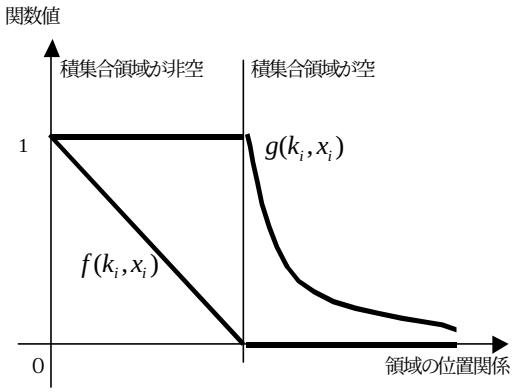

図 3：領域の位置と関数値との関係

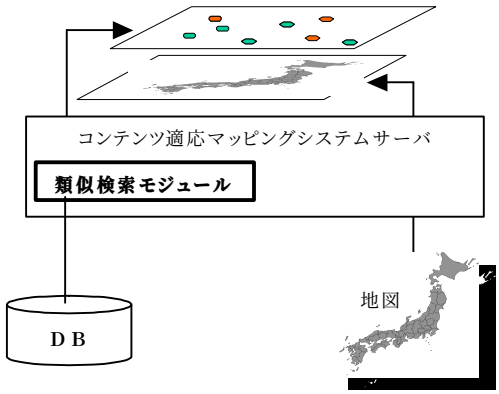

図 4：システムの概観

4. プロトタイプ

3 章で提案した方式について，その有効性を検証するため試作を行った．試作では，筆者らが研究・開発を行っている地図連動型情報提供サービスのコンテンツ適応マッピングシステム[4]の一拡張機能として試作を行った．

コンテンツ適応マッピングシステムとは，統一的な検索インタフェースで地図とDB双方を検索し，DBの検索結果を地図上に重畳表現することを可能にするシステムである．図4で示す様に 3 章で提案した方式を，システムサーバ内の類似検索モジュールとして組み込んだプロトタイプを作成した．この類似検索モジュールによって時空間情報など「域」で表現される属性を含むDBに対して「域」で表現される属性を「点」に縮退させること無く施設を検索キーとして検索する機能を実現した．

5． 評価

本方式の有効性を検証するために従来方式との比較実験を行った．

5.1. 環境

対象データ：横須賀市の飲食店データ 500 件．類似検索に用いる属性は，緯度・経度（点属性），ジャンル（点属性），年齢層（域属性），営業時間（域属性），価格帯（域属性）で，営業時間と価格帯の属性が関連している．

提案手法のパラメタの決定：点属性の緯度・経度に用いる *const* 値の決定の為に予備実験を行った．被験者に複数の施設を提示して「近い」という主観評価の閾値を計測し，その平均値から 350m とする．各属性の重み α の決定は，施設情報DBに対する典型的な検索要求が「場所がどこの施設である」，「いつ営業している」，「価格がいくらである」であることを考慮して，緯度，経度，営業時間，価格帯に対して重みを大きく α=2 とし，他の属性は α=1 とした．また，各属性で線形和を取るときに用いる β は予備実験から決定した．図5に示す予備実験の結果より，ランキング上位 20 件の平均適合率の最大値を取る β=0.7 を採用する．

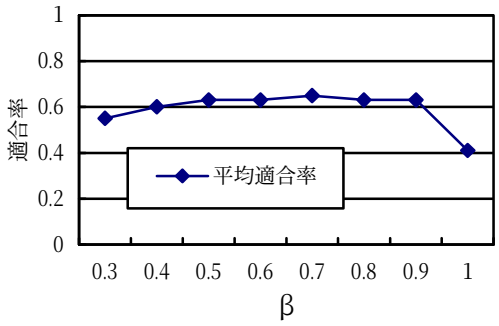

図 5：パラメタ決定の予備実験結果

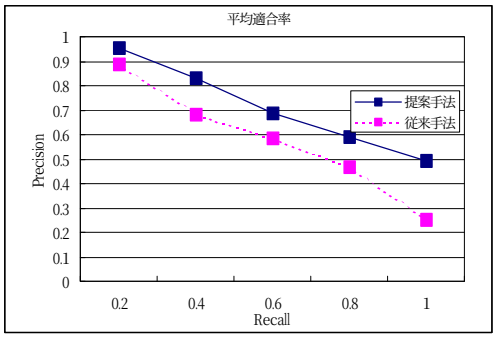

図 6：平均適合率の比較

従来方法:「点」で表現される属性はそのまま「点」で表現される属性として扱い，「域」で表現される属性に対しては代表点を用いた．代表点の算出方法は画像類似検索でオブジェクト位置を特定する一般的な手法[5]である外接直方体の重心を採用した．類似度の計算方法は点間距離に基づく手法を採用した．

評価方法: 異なる検索キーで類似検索を 20 回行い，0.2 刻みの再現率に対する平均適合率を求め，評価指標とした．この方法では再現率が 0.2,0.4 と低い場合の適合率がランキング上位，0.8,1.0 と高い場合の適合率がランキング下位での検索精度を示すものと解釈することが出来る．

正解集合: 適合率判定に用いる正解集合は，3 人の判定者のうち 2 人が正解と判断したものを正解として作成した．

5.2. 実験結果

提案方法と従来方法の比較実験の結果を以下に示す．

図 6 から，提案手法がランキング上位においてもランキング下位においても平均適合率で従来手法を上回っており，提案手法の検索精度の優位性を示している．

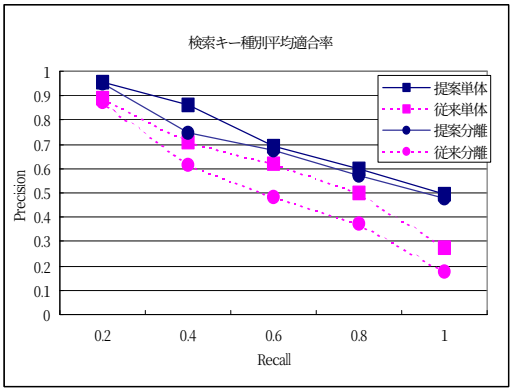

図 7：検索キー種別平均適合率

図7に検索キーが分割・従属表現を含まない場合（単体表現と呼ぶ）と分割・従属表現を含む場合とを区別した検索キー種別の平均適合率を示す．従来手法においては単体表現の検索キーに比べて分割・従属表現を含む方が大きく下回っている．それに対して提案手法では分割・従属表現を含む検索キーに対しても大きく下回ることは無い．よって提案手法は分割・従属表現を含むオブジェクトを検索キーとする場合に従来方法よりも特に優れた方法だといえる．これは提案手法が領域同士の比較を行うことによって各属性の特徴量の性質を失うことなく類似検索に反映させていることに起因するものと考えられる．

6. 関連研究

データベースの検索・インデクシング等の研究において，「域」で表現される属性を「域」として扱った研究が見られる．それらの研究について説明し，本研究との差異を述べる．

時間に関連するデータを検索する研究として，時区間データモデルの研究[6]や TSQL2[7]がある．両者とも「域」で表現されたデータに対する共通演算や被覆演算を定義して「1 月に在籍していた従業員の人数」などを検索可能としている．しかし本研究での従属表現に相当する，時間属性と連携した他の属性というような属性表現には対応しておらず，類似度計算の尺度も用意されていない．

時空間データのインデクシング手法の研究に Segment R-Tree[8]，TP-Index[9]，位相空間データモデル Hawks[10][11]の研究等がある．これらの研究は時間の幅等の特徴量を Hawks では外接直方体の2端点で表現し，TP-Index では開始時刻と終了時刻までの時間という2次元空間上の点で表現してインデクシングを行っている．また演算方法として Hawksでは近似した外接直方体同士を扱う集合演算子として，位相空間の和（union）や差（difference）や積（intersection）等の位相空間演算子が提案されている．しかし特徴量をそのまま扱うのではなく近似している点，類似計算の尺度が無い点が提案手法と異なる．

7. おわりに

本稿では従来最も一般的に用いられてきた，「点」に特徴量を縮退させて最近傍点探索に基づく類似度計算を行う方法ではなく，「点」で表現される属性と「域」で表現される属性に対して各属性の特徴量を領域の集合として管理して，この領域群同士を比較することで類似度を計算する手法を提案し，評価実験を行った．

評価実験では，提案手法の検索精度が従来方法より優れ，特に特徴量を「域」で表現することによって可能になる属性の「分割表現」や「従属表現」の性質を持つオブジェクトを検索キーとする場合に，提案手法が優れていることが明らかになった．この様な多様な属性表現の例として，本研究で用いた「年齢層」，「営業時間」，「価格帯」の他に「スペクトル域」「周波数帯」等が考えられ，提案手法は一定の適用領域を有するものと考えられる．

今後の課題として，本方法では予備実験にて決定していたパラメタ群の効果的な決定支援方法の研究が挙げられる．

参考文献

[1] 串間和彦，赤間浩樹，紺谷精一，木本晴夫，山室雅司：オブジェクトに基づく高速画像検索システム：ExSight，情報処理学会論文誌，Vol.40，No.2，pp.732—741(1999).

[2] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., and Yanker, P.: Query by image and video content: the QBIC system, IEEE Computer, Vol.28, No.9, pp.23—32(1995).

[3] Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D. and Equitz, W.: Efficient and Effective Querying by Image Content, Journal of Intelligent Information Systems, Vol.3, pp.231—262(1994).

[4] 北角智洋，田辺弘実，池田哲夫，星隆司：分散環境における地図情報と連携した情報共有の一手法，第 6 回オブジェクト指向 GIS ワークショップ(2001).

[5] 西村剛，寺本純司，長田秀信，紺谷精一：3 次元物体データベースにおける類似物体検索の検討，情報処理学会 DBS 研究会資料，Vol122，

No.62, pp479—486(2000).
[6] 天笠俊之, 田頭利規, 金森吉成, 増永良文: 時区間データモデルと TSQL2 のデータ選択能力の比較, 情報処理学会 DBS 研究会資料, Vol.109, No.24, pp.141—146(1996).
[7] Snodgrass, R. T.: The TSQL2 Temporal Query Language, Kluwer Academic(1995).
[8] Kolovson, C. P. and Stonebraker, M.: Segment Indexes: Dynamic Indexing Techniques for Multi-Dimensional Interval Data, Proc. 1991 ACM SIGMOD, pp.138—281(1991).
[9] Shen, H., Ooi, B. C. and Lu, H. J.: The TP-Index: A Dynamic and Efficient Indexing Mechanism for Temporal Databases, 10th International Conference on DATA ENGINEERING, pp.274—281(1994).
[10] 黒木進, 牧之内顕文: 位相空間データモデル Universe での空間, 時間, 時空間データ表現, 情報処理学会論文誌, Vol.40, No.5, pp.2404—2416(1999).
[11] 堀之口浩征, 黒木進, 牧之内顕文: 時空間データベースインデックス正規化 R*-tree の実装と性能テスト, 情報処理学会論文誌, Vol.40, No.3, pp.1225—1235(1999).