様式 C-19

科学研究費補助金研究成果報告書

平成２２年５月２８日現在

研究種目：基盤研究（C）
研究期間：2007 ～ 2009
課題番号：19500125
研究課題名（和文） 正例に基づくパターン推論とその知識発見への応用に関する研究

研究課題名（英文） Study on Pattern Inference based on Positive Examples and its Application to Knowledge Discovery

研究代表者
篠原 武（SHINOHARA TAKESHI）
九州工業大学・大学院情報工学研究院・教授
研究者番号：60154225

研究成果の概要（和文）：本研究では，知識発見の対象としてのパターン推論を取り扱った．パターン推論の基礎として，事例データベースからの類似事例の近似検索に注目した．有益な知識発見のためには，大規模な高次元のデータベースからの効率的な近似検索の実現が必要である．本研究で得られた主要な結果は，マルチメディア多次元データベースのための空間索引の高速化技法の開発とそれらを適用した動画同定やプログラム照合への応用実験に集約できる．

研究成果の概要（英文）：This study has been dealt with pattern inference as knowledge discovery. We focus our attention on approximate retrieval of similar examples from database. For useful knowledge discovery, it is important to realize efficient approximate retrieval from large scale high-dimensional database. Results obtained in this study include development of speed-up techniques of spatial index for multimedia high-dimensional database and their application to experimental systems, such as video identification and program matching.

交付決定額

（金額単位：円）

	直接経費	間接経費	合 計
2007 年度	1,200,000	360,000	1,560,000
2008 年度	1,300,000	390,000	1,690,000
2009 年度	800,000	240,000	1,040,000
年度			
年度			
総 計	3,300,000	990,000	4,290,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：学習と発見

１．研究開始当初の背景

(1) 本研究を開始するまでは，複数の事例に共通するパターンを発見する手法に注目していた．逆に，一つの事例に類似する経験事例を検索することも，さまざまな推論の基礎となるものである．従来は，検索のキーに対して完全一致するものを検索することが主流であったが，キーと類似するものを検索す

る場合には，キーそのものをパターンとみなしていると考えられる．こうした類似事例の検索がとくに重要となるのはデータの次元が高次元であり，しかもデータが大量に存在する場合である．

(2) 一方，多次元データのための索引技法はさまざまに考案されていた．索引を効率化するためには，みかけの次元を減らすことにより，いわゆる「次元の呪い」を緩和することが必要である．そうした次元縮小法としては，主成分分析法が有名であるが，そうした技法のほとんどは非ユークリッド空間には適用できないものであった．非ユークリッド距離にも適用可能な次元縮小法としては，代表者らによる Hyper-Map や Simple-Map が開発されていた．

(3) 代表者らは，空間索引を適用するための大規模の多次元データベースとして，動画データから取り出した 1000 万枚を超える画像データベースや音楽データから取り出した 2000 万件の音フレームデータベースを準備して，動画同定や楽曲判定などの応用実験を行っていた．その結果，次元縮小法に Simple-Map，空間索引に R-tree を用いることにより，非ユークリッド空間の大量データに対してもある程度効率的な近似検索が実現できることがわかった．しかし，与えられた事例に近似するものがデータベース中に存在しない場合には，その検索効率が悪化するという問題があった．

２．研究の目的

(1) 本研究は，知識発見の対象としてのパターン推論の基礎として，事例データベースからの類似事例の近似検索に注目し，さまざまなマルチメディア多次元データのための効率的な検索技法を開発するとともにその有効性を応用実験によって示すことを目的とする．

(2) マルチメディアデータの内容に基づく検索においては，データが本質的に高次元であるため，ある程度の不一致を許容する近似検索が必然となる．本研究では，データ間の距離を近似性尺度とする近似検索を対象とする．ここでいう距離は，通常のユークリッド距離に限らず，三角不等式を満たす任意の距離を対象とする．そうした非ユークリッド距離としては，マンハッタン距離（ユークリッド距離は，各座標の距離の 2 乗総和の平方根であるのに対し，マンハッタン距離は，各座標の距離の総和である．L1 距離とも呼ばれる）や文字列間の編集距離（ある文字列を別の文字列に変換するための編集操作の最小ステップ数）などがある．多くのマルチメディアデータではむしろ非ユークリッド距離のほうが自然であることが多い．

(3) 高次元データの情報処理技術の基礎として，パターン認識などの研究成果を用いることができる．ただし，従来の研究方向は，大量データの処理を避けるための手法に主眼が置かれており，これは，マルチメディアデータが本質的に高次元のデータであるため，いわゆる「次元の呪い」を避けることが困難であることに起因している． 次元の呪いを解消するための伝統的手法に，主成分分析法による次元縮小があるが，これはデータ間の類似性尺度にユークリッド距離を仮定しているため，汎用性に欠ける．また，大量データを少数データに代表させるための手法に，ベクトル量子化などの方法があるが，量子化誤差による情報欠損の問題が生じる．本研究では，応募者が考案した SimpleMap 法のように，任意の距離に適用できる次元縮小法を基本とする．

(4) 大量データから高速検索を行う技法としては，2 分探索を発展させた木構造 B-Tree に基づくものが主流であるが，データは 1 次元の順序付けが可能でなければならない．高次元データの高速近似検索はそれほど容易ではなく，たとえば，B-Tree を多次元データ用に拡張した R-Tree を用いた場合でもその守備範囲は 10 次元程度までといわれており，次元が高くなると性能が劣化する．このために，データを次元縮小するなどの方法をとる必要がある．

(5) 本研究では，大量の高次元マルチメディアデータを高速に近似検索することを目指し，そのための索引構造および次元縮小法の開発を行い，画像や音声などに対する有効性を示す．索引技法としては，空間索引構造 R-Tree および応募者の考案した次元縮小法 SimpleMap を基本とする．R-tree は，代表的かつ標準的な空間索引構造であるが，これを非ユークリッド距離データへ対応でき，動的構築法が行えるように改良し，さらに効率化を図る．また，SimpleMap は，非ユークリッド距離データを取り扱うことができるという特徴をもつが，射影軸の選択法や文字列データやゲーム盤面データへの適用などについてさらに改良を加え，より効果的なものとする．さらに，音声認識や音楽データ検索，動画同定，バイナリプログラム照合，ゲーム局面検索など，さまざまなデータへの適用実験を行う．また，ごく最近になって新しい検索法として，スケッチ（近似性をある程度保持する一種のハッシング）を用いる方法が注目されているので，これをさらに発展させて，R-Tree による手法と融合を目指す．本

研究の成果は，情報検索への貢献だけでなく，新しい形のパターン認識手法などの創出の可能性も示せるものと期待できる．

３．研究の方法

(1) データベースとしては，画像や音，プログラム断片，将棋盤面などのさまざまな種類の大量データを用いる．これらのデータにおいては，非ユークリッド距離が用いられているので，これらを効率よく検索するための技法を開発する．その有効性を示すための応用実験を行う．

(2) 大量の高次元マルチメディアデータを高速に近似検索することを目指し，そのための索引構造および次元縮小法の開発を行い，画像や音声などに対する有効性を示す．索引構造は，R-Tree を基本として，その構築法の見直しを行う．次元縮小法は，SimpleMap を基本として，これをさまざまなデータに適用できるように拡張するとともにより効果的な射影を求める手法を確立する．マルチメディアデータとしては，静止画や動画，音データ，将棋盤面データ，プログラム断片などを対象とし，それらを索引付けして高速近似検索を実現し，その効果を実証する．

(3) まず，これまの研究の結果を整理し，次元縮小法および R-Tree 構築法の見直しを行うとともに，画像や音声などのデータを索引付けしてその効果を実証する．具体的には，以下の項目を中心に研究を行う．

① 空間索引技法に関する研究

（ｉ）次元縮小法 SimpleMap の改良

射影軸の選択方法について，これまでのランダムサンプリングによるものを基本として，より情報欠損が少ない高能率な射影軸を求めるアルゴリズムを開発する．

（ⅱ）R-Tree 構築法

空間を階層的に分割して索引付けを行う場合，その分割手法が重要である．ヒルベルト空間補填曲線による順序付けが最適であることが知られている．その順序に沿ったソートを高速に行うアルゴリズムを開発する．

（ⅲ）局所次元縮小射影の導入

次元縮小は全データに対して最適なものを選択するのが通常であるが，R-Tree により階層的に分割された部分空間に対して局所的なデータの偏りが生じる可能性がある．この局所的なデータの偏りを反映した射影次元を追加することにより検索の高速化を行う．

② マルチメディアデータの近似検索に関する研究

（ｉ）以下のような具体的データに対する応用実験を行う．いずれの場合においても，データベースが巨大となる場合にも高速に検索できることが重要であるので，本研究で開発する手法がいかに効率的であるかが鍵となる．これらの実験を通じて，本研究で開発する手法の有効性を実証するとともに，さらなる改良点の洗い出しを行う．

（ⅱ）動画同定システムの実現

動画は連続する静止画フレームから構成される．個々のフレームに動画ラベルを付して索引付けするようにする．動画同定システムでは，まず，大量の動画をデータベースに登録しておき，質問として与えられた動画がデータベースに登録されている動画のいずれであるかあるいはいずれでもないかを同定する．動画同定システムは，たとえば，動画サイトなどに違法にアップされた著作権侵害の検出やテレビ CM の放映監視などに応用することが可能である．

（ⅲ）楽曲検索システム

約 2000 曲を音楽 CD から取り出し，その周波数特性などを特徴として検索を行う．類似な曲をさがすというよりは，部分的な音データからもとの曲を判別するというようなものとなる．ただし，検索時にはラジオやテレビ放送，テープ録音などの劣化した音源による検索が可能となることを目指すとともに，複数の楽曲をいろいろな割合で混合した合成曲から原曲を判別する実験も行う．

４．研究成果

(1) 空間索引を高速化する技法としては，従来手法では効率が悪化する類似事例がデータベース内に存在しない質問を高速化する手法として，質問処理の履歴を活用する手法を考案し，その有効性を確認した．

(2) 高速な空間索引の応用例としては，オープンソースソフトウェアの保護のためのライセンス違反検出のためのプログラム照合や著作権侵害の動画アップロード検出のための動画同定システム，複数曲の混合からの原曲判別，棋譜データベースからの類似局面の検索など，さまざまな場面に応用し，その有効性を確認した．

(3) また，新たな空間索引法として，縮小型構造データ Sketch について，これを非ユークリッド距離に適用できるように改良する研究にも着手し，その方向の妥当性を確認した．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

① Developments from enquiries into the learnability of the pattern languages from positive data, Theoretical Computer Science, vol. 397, no. 1 - 3, pp. 150 - 165, (2008), (Yen Kaow Ng, Takeshi Shinohara)，査読有

〔学会発表〕（計 12 件）

① Arnoldo José Müller Molina, Takeshi Shinohara: On the Configuration of the Similarity Search Data Structure D-Index for High Dimensional Objects. Proceedings of the 2010 International Conference on Computational Science and Its Applications (ICCSA 2010) (3), pp.443-457, 査読有, 2010 年 3 月 25 日, 九州産業大学.

② Takaaki Aoki, Daisuke Ninomiya, Arnoldo José Müller Molina, Takeshi Shinohara: Accelerating Video Identification by Skipping Queries with a Compact Metric Cache. Proceedings of the 2010 International Conference on Computational Science and Its Applications (ICCSA 2010) (4), pp.252-262, 査読有, 2010 年 3 月 23 日, 九州産業大学.

③ Arnoldo José Müller Molina, Takeshi Shinohara: Efficient Similarity Search by Reducing I/O with Compressed Sketches. 2009 Second International Workshop on Similarity Search and Application (SISAP 2009): pp.30-38. 査読有, 2009 年 8 月 29 日, プラハ.

④ 岩崎瑶平, Arnoldo Müller, 篠原武: Sketchによる大容量記憶データに対する高速な空間検索法に関する研究, 人工知能学会第 74 回人工知能基本問題研究会 (SIG-FPAI), pp.7-11, 査読無, 2009 年 9 月 14 日, 広島市立大学.

⑤ 青木隆明, 田島圭, 篠原武: R-treeの検索高速化に関する研究～ノードへのデータ配置法の提案～, 人工知能学会第 74 回人工知能基本問題研究会 (SIG-FPAI), pp.13-18, 査読無, 2009 年 9 月 14 日, 広島市立大学.

⑥ 青木隆明, 岩崎瑶平, 篠原武: 棋譜データベースからの近似盤面検索に対する効果的な射影法に関する研究, 火の国情報シンポジウム 2009, B-7-1, 8 pages, 査読無, 2009 年 3 月 14 日, 九州産業大学.

⑦ 浦郷祐希, 田島圭, 青木隆明, 岩崎瑶平, 篠原武: 空間索引による近似画像の高速検索を用いた動画同定システムの実現, 火の国情報シンポジウム 2009, B-7-2, 8 pages, 査読無, 2009 年 3 月 14 日, 九州産業大学.

⑧ 棚町直矢, 松崎陽太, 青木隆明, 岩崎瑶平, 篠原武: 音楽データベースを用いた合成曲の原曲判別システム, 火の国情報シンポジウム 2009, B-7-3, 6 pages, 査読無, 2009 年 3 月 14 日, 九州産業大学.

⑨ 二宮大輔, 今村安伸, 市来亮, 篠原武: 大量質問点の順序付けによる類似検索高速化に関する研究, 火の国シンポジウム 2008, B-3-1, 6-pages, 査読無, 2008 年 3 月 6 日, 長崎大学.

⑩ 市来亮, 今村安伸, 二宮大輔, 篠原武: 動画検索のための画像の正規化と特徴抽出について, 火の国シンポジウム 2008, B-3-3, 6-pages, 査読無, 2008 年 3 月 6 日, 長崎大学.

⑪ 今村安伸, 市来亮, 二宮大輔, 篠原武: 空間索引のための射影法, 火の国シンポジウム 2008, B-3-4, 6-pages, 査読無, 2008 年 3 月 6 日, 長崎大学.

⑫ Arnoldo José Müller Molina, Takeshi Shinohara: Fast Approximate Matching of Programs for Protecting Libre/Open Source Software by Using Spatial Indexes. Seventh IEEE International Working Conference on Source Code Analysis and Manipulation (SCAM2007): pp.111-122. 査読有, 2007 年 9 月 30 日, パリ.

６．研究組織

(1)研究代表者

篠原　武（SHINOHARA TAKESHI）

九州工業大学･大学院情報工学研究院･教授

研究者番号：60154225

(2)研究分担者

なし

(3)連携研究者

なし