汎化冪空間類似度法によるデータパーセプション技術の開発

1. 背景

IT社会の到来により、近年、大量の情報の収集、送受信、処理、蓄積コストが急激に下がった。われわれは以前と比べて多くの情報を利用することができるようになった。その一方でデジタル情報の洪水が発生している。IT社会の現実の姿は、ほとんどの情報は蓄積されるのみで分析に活用されることがない。このような背景から、産業分野や学術分野において、大量のデータを高速に分析し、推定・予測を行える技術が必要とされるようになってきている。しかしながら、従来の確立されたデータ分析手法の多くは、データ件数、カテゴリー数に対してスケーラブルでないものが多い。そのため、大量の情報を前にして、現在のITは時間計算量がネックとなってインテリジェンスの欠如を起こしている。

一方、「ムーアの法則」、「ストレージの法則」と呼ばれる言葉が示すとおり、計算機の性能向上、特に、メモリ容量、各種ストレージ容量の増大は目覚しいものがある。このような大量の記憶空間を効果的に使用することに長けた手法がITのボトルネックを解消する有望なシーズであり、新しい技術スキームを創出する可能性を持つものと、提案者は従来より着目してきている。

2. 目的

このような背景から、本開発では、大きな記憶空間を使用することによって高速計算を実現し、大量のデータ件数、カテゴリー数に対してもリアルタイムに処理できるソフトウェア基盤技術を実装することを目的とする。それにより、従来は扱うことの出来なかった領域で有効な作用が得られ、データパーセプション・データインテリジェンスと言うべき、新しいフレームを創出することが期待される。

提案者が過去に開発した冪空間類似度法は、学習が非常に高速で、かつ、認識も非常に高速に行えるパターン学習・認識アルゴリズムである。これを拡張した「汎化冪空間類似度法」を用いて、大量のデータ件数、カテゴリー数に対してもリアルタイムに処理できるソフトウェア基盤技術の開発を行う。本技術の主要なアイデアを簡単に言えば、「時間計算量を空間計算量に転換する」ことにより大量のデータを高速に分析し、知的な処理を行うというものである。

3. 開発の内容

本プロジェクトでは、汎化冪空間類似度法を用いたさまざまなデータ分析技術のプログラムによる実装を行った。開発リソースの枠組みの中で、基礎的で汎用性のある基盤プログラムの開発に範囲を絞った。具体的には、最近傍探索プログラム、パターン学習・認識プログラム、カテゴリーの相関分析プログラム、コンテンツの自動抽出システムの開発である。

3.1 最近傍探索プログラム

近似最近傍探索のための前処理を行うプログラムと、探索を行うプログラム。

Cのライブラリ関数の形式で提供される。また、サンプル用のドライバ関数を作成した。性能評価を行ったところ、同一の検索精度で速度比較を行い、ある条件では線形検索の8

倍程度の高速化が達成できた。また、別の条件では10～20倍程度の高速化が達成できた。本プログラムは、検索精度と検索時間を動的に選択できるという特長がある。速度と精度の関係を計測したグラフを図1に示す。

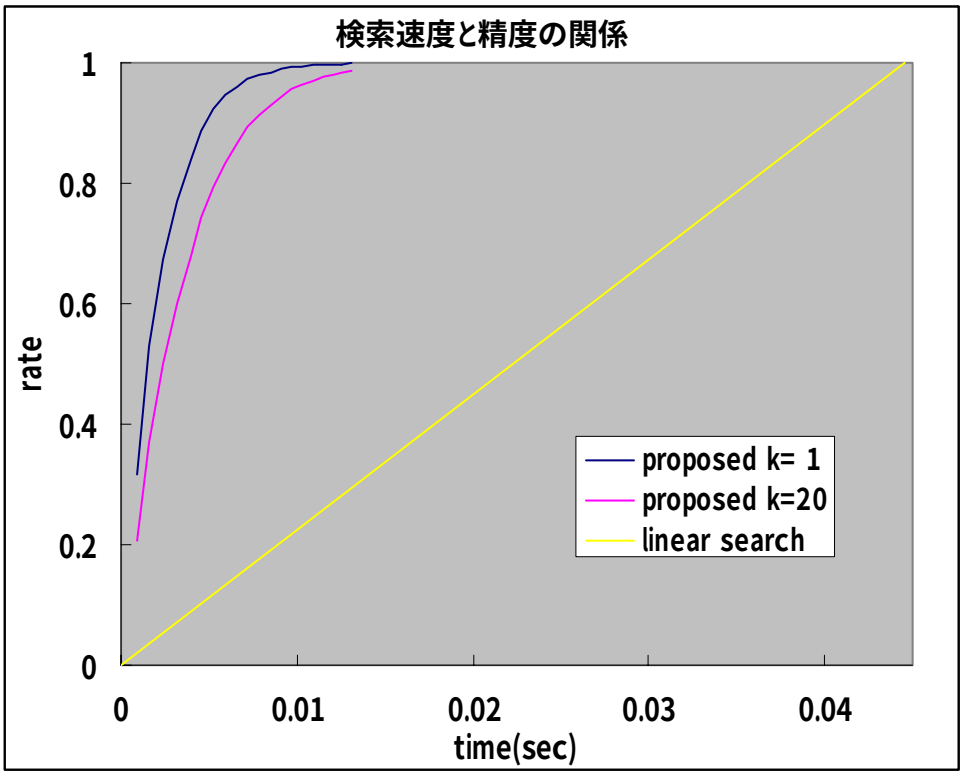

図1　検索速度と精度の関係（20次元10万データの場合）

3．2　パターン学習･認識プログラム

高速な学習･認識プログラムを作成した。本プログラムは、学習サンプル数、カテゴリー数に対してスケーラブルである。本プログラムでは仕様上、学習サンプル数、カテゴリー数は $2^{31}-1$ までと、事実上無制限である。

Cのライブラリ関数の形式で提供される。また、サンプル用のドライバ関数を作成した。公開されているオフライン手書き数字データベース MNIST を使った実験では、認識率95.94%であった。認識速度は、k-NN 法の 700 倍以上高速だった。

また、ソースコードは、文字認識に特化したものではなく、パターン認識全般に適用できるように設計されている。

3．3　カテゴリーの相関を分析するプログラム

カテゴリー間の距離の定義のしかたはいろいろ考えられるが、いずれにしても、それぞれのカテゴリーについて近いカテゴリーを検索するとき、全カテゴリーについて処理を行う場合、カテゴリー数に対して2乗の時間がかかるし、サンプル数に対しても2乗の時間がかかると予想される。本プログラムでは独自のアルゴリズムを使用して、これを高速に行うことができる。

例えば、活字ＯＣＲ用データベースを使って、日本語全ての文字種について形の似ている文字の上位 100 個を求める処理を行ったところ、文字種（カテゴリー数）＝6891、サンプル数＝440621 に対して、２０分程度で処理を行えた。

分析結果のサンプルを図2に示す。

この手法は、データマイニング、ウェブマイニングなどに応用が可能であると考える。

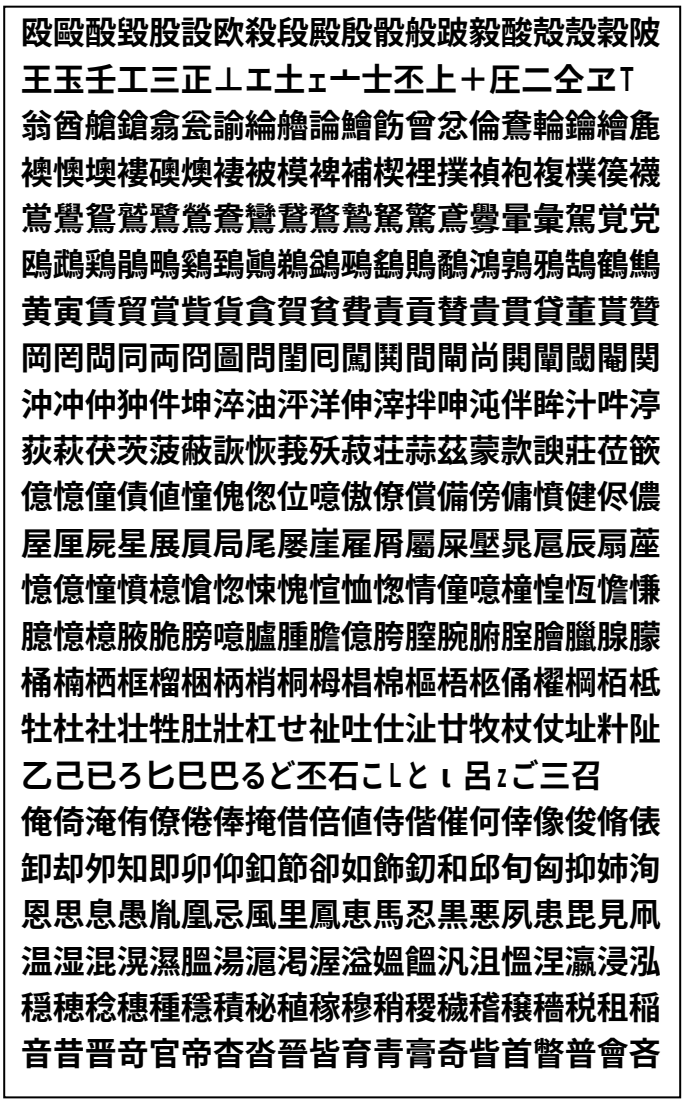

図2　活字日本語文字種全てについて形の似ている順位で並べた結果

3．4　点の均等配置プログラム

点を決められた領域内に均等に配置する点の均等配置問題は、一般に多くの計算時間を必要とする。3次元球面の点の均等配置の例を図3に示す。

本プロジェクトにおいて、独自のアルゴリズムを用い、高次元球面の点の均等配置プログラムを作成した。10 次元、1024 点の条件で実行し、高速に行えることが確認できた。

本プログラムの応用により、科学技術分野の計算処理が高速に行え、科学的な新しい知見の獲得のためのツールとなることが期待される。

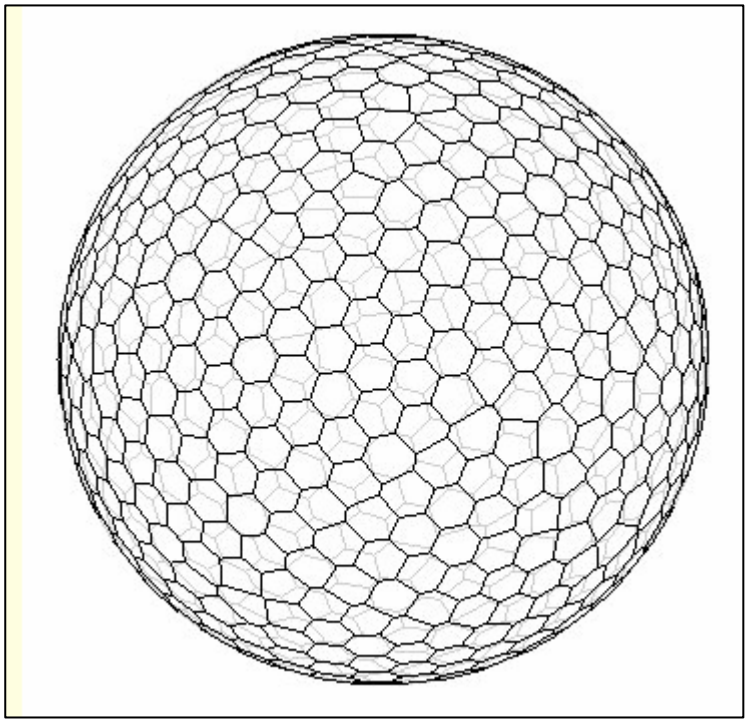

図3　点の均等配置のシミュレーション結果 (球面上の 701 個の点の場合)

3．5　コンテンツの自動抽出システム

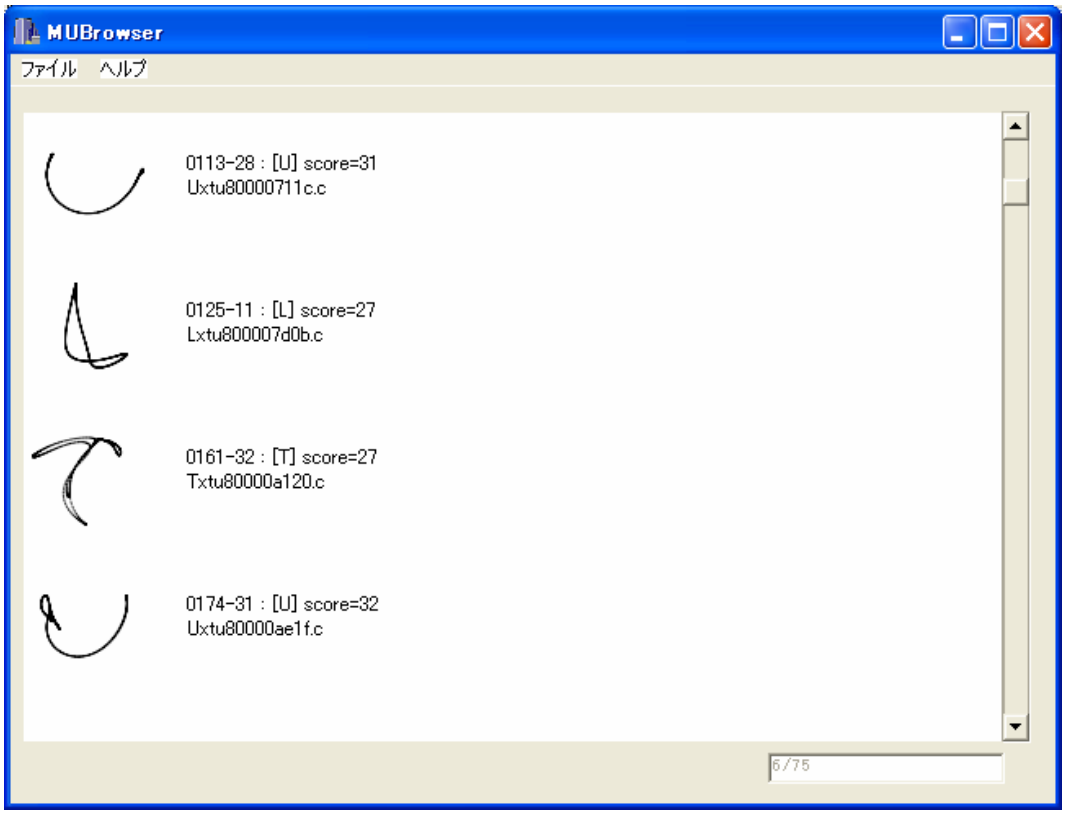

図4　コンテンツアーカイブ登録ツールの画面

たくさんの画像の中から、ある目的に沿って、特定の画像を抽出するシステムのプロトタイプとして、本技術を応用した、コンテンツの自動抽出システムを作成した。

システムの概要を以下に説明する。

適当な画像生成式とランダムなパラメータにより、ランダムな画像をプログラムにより多数生成する。文字認識プログラムを使って、こうしてできたランダムな画像の中から、文字に似ている画像だけを抽出する。それらのうち、人間が画像を確認して、意味を感じるもののみ、アーカイブに保存する。この一連の作業により、意味のある画像とそれを生成するプログラムを多数獲得することができる。こうして得られた画像・プログラムは、全く新しい方法で作成されたコンテンツである。従って、新しい利用方法が生み出される可能性がある。

このプロトタイプシステムを作成し、テスト作業を行ったところ、有効なコンテンツを得られることが確認できた。

図4は、必要な画像をアーカイブに保存するための、人間の補助となるアプリケーションプログラムの画面である。

4．従来の技術（または機能）との相違

従来のデータ分析技術のほとんどは、データ数に対してスケーラブルでないものが多い。本技術はスケーラブルな性質を持っており、データの前処理と探索の両方の時間計算量を小さくすることが可能である。従って、非常に大量のデータの分析に強みを持つ。

5．期待される効果

本技術は主に次のような分野に適用が期待できる。

・情報探索、推定・予測、リアルタイム意思決定、など
・探索がボトルネックとなっているシステム全般（ビジネス・学術）
・大量のデータ同士の相関を分析することにより実現されるようなもの
　（ex．ウェブページ同士の関係を分析することによる検索エンジンの作成）

6．普及（または活用）の見通し

現在、営業活動を行っている。

7．開発者名（所属）

小林卓夫（東京農工大学工学教育部博士後期課程）

（参考）開発者ＵＲＬ

http://www.geocities.jp/onex_lab/