Data Rearrangement Engine

データマイニングのための データ再配列エンジン

国立情報学研究所
北本　朝展
http://research.nii.ac.jp/~kitamoto/

本論文の基本的立場

1. データマイニングとは，データベース（データ集合）からデータに関する興味深い性質を発見するための手法である．
2. 決して全自動プロセスではない．
3. 人間がデータベースと対話しながら，今まで気付かなかったデータの性質を発見していくプロセスである．
4. ゆえに，データベースと対話するための強力な問い合わせ言語が必要である．

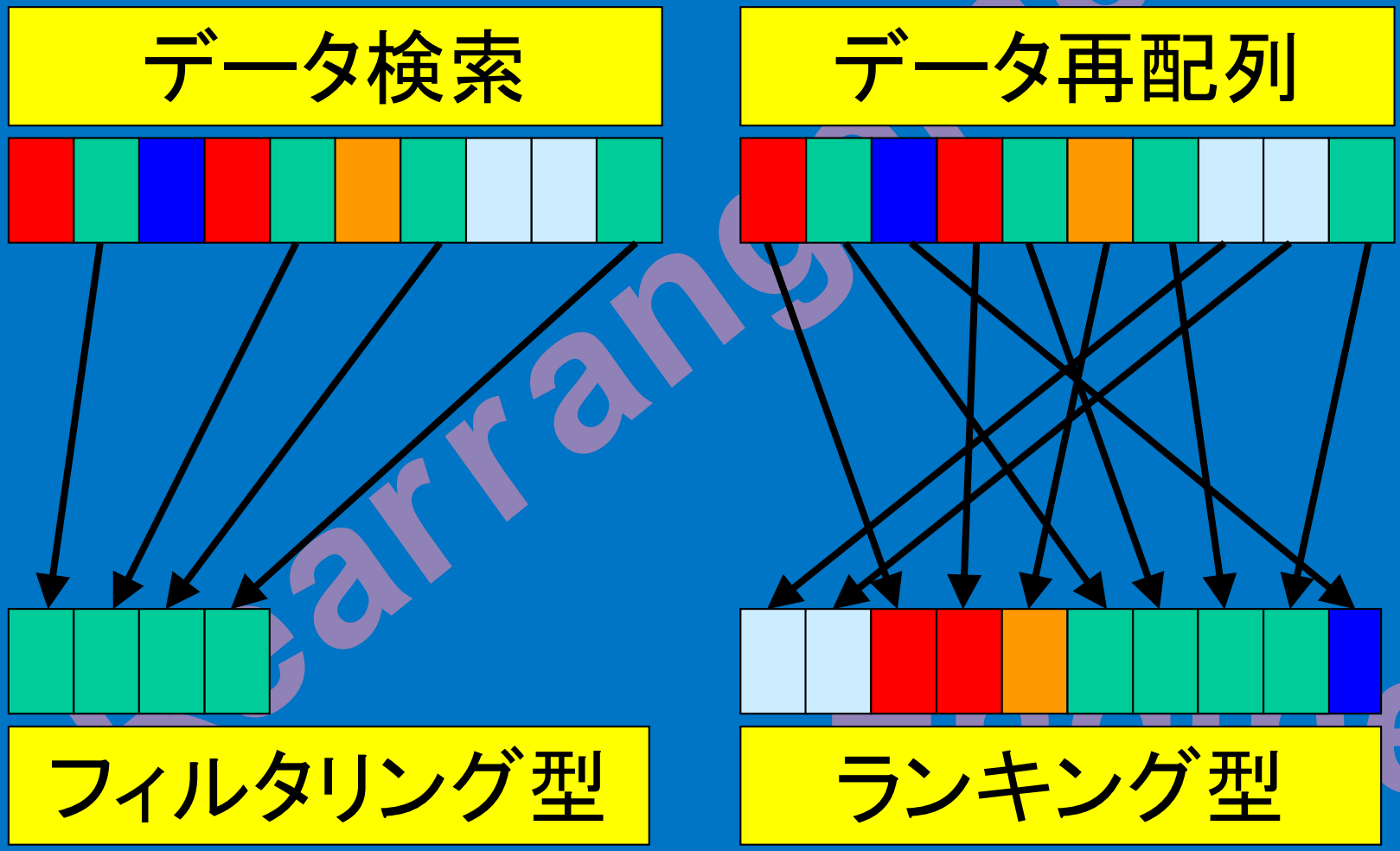
問合せ言語

- データベースエンジンに対し関心あるデータを特定するための条件を伝える言語.
- データベース分野では伝統的な研究テーマであり長年にわたる多数の研究がある.
- 実用的にはSQL (Structured Query Language)が圧倒的優位を誇る.
- しかしデータマイニングのためのデータ操作のモデルを考え直す必要はないか？

本発表の概要

1. はじめに
2. データ検索とデータ再配列
3. データ再配列エンジンの問い合わせ言語における基本的演算子
4. データ再配列の具体例
5. 「デジタル台風」における利用
6. おわりに

データ検索とデータ再配列



データ検索(Retrieval)

- 関心のあるデータのみをデータベースから引き出すことを目的とする，フィルタリング型のデータ操作．
- 関心のあるデータを明示できる場合に，データの絞込みが効率的におこなえる．
- これまでのデータベース研究の多くはこの種の問い合わせについて研究してきた．
- これだけで十分なのか？

データ再配列（Rearrangement）

- 重要なデータを先頭に配置し，類似したデータを近くに配置すること（再配列）を目的とする，ランキング型のデータ操作．
- 適切な並べ替えによりデータの関係に着目し比較することが可能となる。
- データマイニングの場合には，こちらのデータ操作の方が有用ではないか？
- この仮説に基づく問い合わせ言語の構築．

データ再配列の操作

- 情報検索や画像検索において一般的なランキング・類似検索はその代表的な操作.
- データをまとめるグルーピングも，一種の並べ替え操作とみなせる.
- その他，データのクラスタリングや分類も，データ再配列操作の一種として理解できる.
- 上記の操作をデータ再配列として統一的に記述可能な問い合わせ言語の構築。

SQL

- 何を対象とし(FROM)，どのような条件を満たすもののうち(WHERE)，この属性を示せ(SELECT)，という形式で記述する．
- 集合操作を基本とする関係代数，述語論理を基本とする関係論理，などの数学的モデルを背景としている．
- しかし数学的モデルそのものではなく，種々の実用的改良を施している．

関係代数の演算子（1）

- 5個の基本的リレーショナル代数演算子
  1. 和（union）
  2. 差（difference）
  3. 直積（Cartesian product）
  4. 射影（projection）
  5. 選択（selection）
- 集合に対する演算の結果が集合になるという閉じた体系であることが重要.

関係代数の演算子(2)

- 関係代数の真価は，実用的に重要な以下の演算子が，基本的演算子から導出可能であることを示したこと.
  1. 結合(join)，特に自然結合(natural join)
  2. 共通部分(intersection)
  3. 商(division)
- これらを複雑に組み合わせても破綻しないため，複雑な問い合わせを記述可能.

関係代数とSQL

- SQLは関係代数に基づく言語であるが，関係代数を実用的に改良もしている．
- 例えばグループ化演算子（GROUP BY）や整列演算子（ORDER BY），集約演算子(Aggregation)等を追加した．

- この改良の含意を深く考える必要がある．
- これらの演算子は本質的に重要かも？

SQL改良部分と情報検索（1）

- グループ化演算子
  - クラスタリングに対応している.
  - 類似したデータをまとめて操作対象のデータ数を減らすために重要なデータ操作である.
- 整列演算子
  - ランキング，類似検索などに対応している.
  - 重要なものを先頭に，類似したものを近くに配置するために重要なデータ操作である.

SQL改良部分と情報検索（2）

- 集約演算子
  - データ要約に対応している.
  - 集合の統計量や特徴を簡潔に記述するために重要なデータ操作である.
- 最近の研究では，このような改良部分に関連する部分の研究が盛んである.
- 実はこれらの部分こそ，データマイニングにとって本質的に重要なのではないか？

新しい問合せ言語

- 上記の操作を総称する言葉として「再配列（rearrangement）」という言葉を用いる.
- その意図は，「*データベースというデータ集合に対し，その要素をユーザの要求に応じて配置しなおして提示する*」こと.
- 従来型の問い合わせ言語とは最終目的が異なるため，SQLのつぎはぎ拡張ではなく，新しい問合せ言語が必要である.

既存の問合せ言語との比較

- SQL/MM　マルチメディア用途
  - SQLのMMへの拡張であるが、データ検索というパラダイムは強固に残っている.
- DMQL　データマイニング用途
  - SQLとの統合のしやすさを最優先にしつつ，一定のデータマイニング操作をサポート.
- MRML　画像検索用途
  - 基盤となるモデルが明瞭ではなく，基本構造や拡張は系統的ではない.

本発表の概要

1. はじめに
2. データ検索とデータ再配列
3. データ再配列エンジンの問い合わせ言語における基本的演算子
4. データ再配列の具体例
5. 「デジタル台風」における利用
6. おわりに

データベースのモデル

- オブジェクト関係モデルに近いモデル.
- 論理構造は表形式であるが，属性は内部構造をもってもよい（例えば，リスト，ベクトル，XML構文木などの複合オブジェクト）.
- 属性の型に応じた演算子を用意する．例えば，型による類似尺度の定義の相違は，ここで吸収する.

オブジェクトリレーショナルモデル

	属性1	属性2	属性3
タプル1	1000	TOKYO	XML1
タプル2	2000	SAPPORO	XML2
タプル3	2500	OSAKA	XML3

表の形式で表現するが、各属性は内部構造をもつ複合オブジェクトであってもよい。

基本的な演算子

1. グループ化（grouping）
2. 整列(ordering)
3. 属性拡大（attribute expansion）
4. 収集(gathering)

これらの演算子の組み合わせで，各種のデータ再配列操作を記述する．

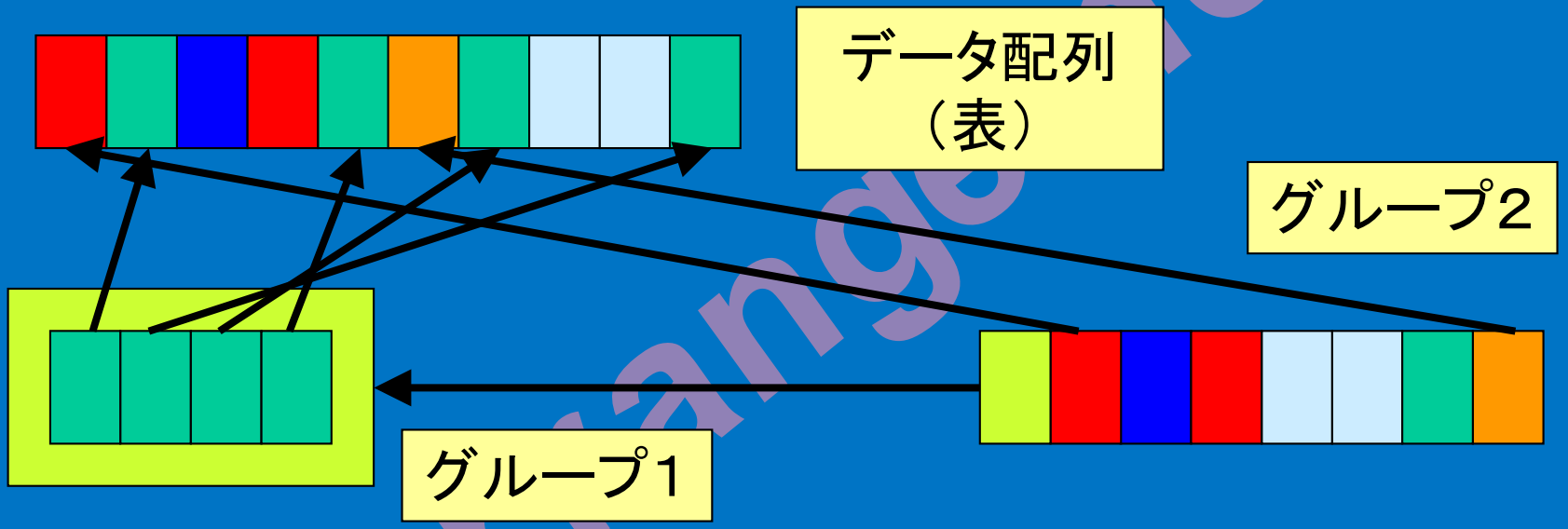
グループ化（grouping）

- ある基準のもとで一まとまりとみなせる要素の集合を「グループ」と呼ぶ.
- このようなグループを新たに生成する演算子をグループ化とよぶ.
- データを「要素」と定義するが，新たに生成したグループも「要素」と定義する.
- すると要素のグループが再びグループとなるような閉じた体系を考えることが可能.

グループ化の例

- ある属性をキーとして同一の属性値をもつデータをまとめる操作.
- 統計学における層別化のように，ある属性の定義域をいくつかの区間に離散化してデータをまとめる操作.
- 類似したデータをまとめる階層的あるいは非階層的クラスタリング操作.

物理層から見たグループ化



- 配列は順序つき配列「リスト」，その各要素には特定データを一意に指定するための論理的情報「ポインタ」が存在する．

整列(ordering)

- 整列とはグループ内の要素を大小関係にしたがって並べ替えること.
- 大小関係の基準は，永続的属性あるいは一時的属性のどちらでもよい.
- ランキング検索，類似検索などにおいて重要なデータ操作.
- 情報検索用途では、大小関係が問い合わせに依存して決まることも多い。

属性拡大(attribute expansion)

- 属性には以下の2種類を考える
  1. 永続的(persistent)属性
  2. 一時的(volatile)属性
- 問合せを処理する間にのみ存在する属性を一時的属性とよぶ.
- 属性(ほとんどは一時的属性)を拡大するために, この演算子を用いる.

収集（gathering）

- グループから要素を選択して収集し，新たなグループを生成する演算子.
- これは，ランキング検索において，上位X件を選ぶというような操作に対応する.
- 他の演算子によって表現可能であるため，独立した基本演算子とは言えない.
- この演算子は頻出の演算子であり，問い合わせ最適化にも寄与する.

問い合わせ最適化

- ところで、データが1億件でも全部を整列するのか？それは非現実的ではないか？
- 上記の演算子に関する記述は，あくまで論理層での記述である.
- 物理層においては，もっと効率的な「実行プラン」を生成できる. 例えば上位X件の要素しか必要ないなら，その性質を活用した問い合わせ最適化が可能である.

Data 要素の配置

- 「配置」というのは，要素の並びの論理的構造を意味する.
- 要素の配置は，一次元空間にとどまらず，高次元空間での配置も一般的である.
- その視覚的レンダリングについては可視化インタフェースが担当する.
- データベースエンジンと可視化インタフェースとの協調が必要である。

本発表の概要

1. はじめに
2. データ検索とデータ再配列
3. データ再配列エンジンの問い合わせ言語における基本的演算子
4. データ再配列の具体例
5. 「デジタル台風」における利用
6. おわりに

Data 類似画像検索

1. 各要素に対して「距離」という一時的属性を拡張する「属性拡張」演算子を適用。
2. 属性値の計算法として例示画との距離尺度を定義し、各要素の属性値を計算。
3. 「整列」演算子を適用して、「距離」属性の昇順に整列。
4. 「収集」演算子で上位N件の要素を収集し、そのグループを検索結果として応答。

時系列画像類似検索

1. 「グループ化」演算子を適用して、時系列ごとに画像をグループ化する。
2. グループごとに時間を基準としてデータを整列する。
3. 各グループごとに距離属性を拡張し、属性値の計算法として例示系列とのDynamic Time Warpingを指定し、これを用いて距離の属性値を計算する。
4. 「グループ化」演算子を適用し、時系列ごとの距離を属性値として含む配列を生成する。
5. 「整列」演算子を適用し、距離属性の昇順に整列する。
6. 「収集」演算子で上位N件の時系列を収集し、そのグループを検索結果として応答する。

クラスタ代表画像検索

- クラスタ代表点を計算する。その方法として、例えばK-means法や自己組織化マップなどを指定する。
- クラスタ代表点とデータの距離が最小となるグループに、すべてのデータを「グループ化」する。
- クラスタ代表点とデータとの間の距離を一時属性「距離」に記録する。
- 整列演算子を各グループに適用し、一時属性「距離」の昇順に整列する。
- 「収集」演算子を適用し、各グループから1件を収集する。
- これらの要素を新たなグループとし、その際にはクラスタの配置に関する情報を付加して応答する。

データ分類

- データ分類に必要なパラメータはあらかじめセットするか、学習する。
- 表を「スコア」という一時属性で拡張し、その属性値を固有ベクトル上への射影関数で計算する。
- 「グループ化」演算子を適用し、しきい値との比較の正負でデータを2グループに分割する。

具体例のまとめ

- 基本的な演算子の組み合わせによって多様な計算が可能となることを示した。
- これらの例はいずれも他の方法によって実現可能ではある。
- しかし、代数的な演算の導入によって、処理の自由度は大幅に向上する。
- 特に入れ子的な問い合わせに対する効果が大きい。

本発表の概要

1. はじめに
2. データ検索とデータ再配列
3. データ再配列エンジンの問い合わせ言語における基本的演算子
4. データ再配列の具体例
5. 「デジタル台風」における利用
6. おわりに

デジタル台風における利用（1）

- 気象衛星画像から生成した台風画像のコレクション。現在の画像数は46000件程度であり、これをWWWで提供している。
- 現在の事例に類似した過去の事例に基づき、事例ベースの意思決定を支援する。
- 現在のアクセス数は1日500件程度。ただし台風が日本に接近すると、1日20000件程度に増加することもある。

デジタル台風における利用(2)

- 問い合わせ言語はXML構文で記述。
- デモンストレーション
  1. 類似画像検索
  2. クラスタ代表画像検索
- 他にもデータの関係を探り、データ間の比較を可能とするための機能を実装中。

http://www.digital-typhoon.org/

何のための類似画像検索？

- ある画像に類似した画像が探せると，何が嬉しいのか？
  1. 類似したデータに見られる何らかの共通性・差異性を見るための手段.
  2. 最終的に欲しいデータにたどりつくためのデータベース探索の一つの手段.
- 本当の用途を見定めたうえで、それに必要な操作を定めていく必要がある。

まとめ

1. 「データ再配列」という考え方がデータマイニングにおいて有用であると主張した。
2. データ再配列において基本的な演算子を提案し、それを用いた問い合わせ言語を提案した。
3. この言語を用いると、類似（時系列）検索やクラスタリング、分類などが記述可能であることを示した。

今後の課題

1. 本論文で提案したモデルを、よりフォーマルに定式化する必要がある。
2. 上記の問い合わせ言語を完全にサポートするエンジンを実装する。
3. 実世界問題において、このようなデータ操作が実際に有用であることを示す。
4. エンジン高速化のための問い合わせ最適化などについて議論する。