新技術説明会
New Technology Presentation Meetings!

# データ量の大幅圧縮と検索速度向上を実現するデータベースの構成法

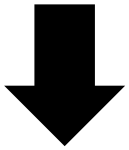
神奈川大学　大学院理学研究科　理学専攻
情報科学領域
教授　桑原 恒夫

2018年12月4日

# 研究分野の背景

ビックデータの時代

大量のデータ蓄積とその利用による様々な価値創出



保存するデータ量の圧縮や検索速度の向上は重要な課題

# 従来技術とその問題点

現在のデータベースマネージメントシステム(DMBS)の主流は
リレーショナルデータベース(RDB)

RDBの検索速度向上のための主な手法はインデックス法*

しかし

インデックス法では、検索速度は向上するがデータ量は減少しない
（むしろインデックスをつける分、データ量が若干増加する）

*データベース記録した内容に索引をつけ、検索時にその索引から辿る。記録の全部を参照しなくてもよくなり、検索が高速に実行できる。しかしその索引のデータは余分に記録する必要がある。

# 新技術の特徴・従来技術との比較

- 2つのカラム間の関係を、新たに導入した上位項目という概念を利用して記録

⬇

データ量を大幅に圧縮できる。

- 検索すべきデータ量が圧縮されるので、検索速度を向上できる。
- データの正規性（冗長でないこと）、完全性（データの欠損がないこと）は保持できる。
  （したがって従来形式のデータ構造に戻すことも可能）
- 但し、データの更新（入力）には従来より時間がかかる。

# 新技術の基本原理

新技術説明会
New Technology Presentation Meetings!

## 従来法

| 項目A | 項目B |
|---|---|
| a | 1 |
| | ・ |
| | ・ |
| | 500 |
| | 501 |
| | ・ |
| | ・ |
| | 1000 |
| b | 501 |
| | ・ |
| | ・ |
| | 1000 |
| | 1001 |
| | ・ |
| | ・ |
| | 1500 |
| c | 1001 |
| | ・ |
| | ・ |
| | 1500 |
| | 1501 |
| | ・ |
| | ・ |
| | 2000 |

合計3000レコード

## 提案方法

| 項目B | 項目Aか上位項目 |
|---|---|
| 1 | a |
| ・ | |
| ・ | |
| 500 | |
| 501 | X |
| ・ | |
| ・ | |
| 1000 | |
| 1001 | Y |
| ・ | |
| ・ | |
| 1500 | |
| 1501 | c |
| ・ | |
| ・ | |
| 2000 | |

2000レコード

**上記項目の定義**

| 上位項目 | 項目 A |
|---|---|
| X | a |
| | b |
| Y | b |
| | c |

4レコード

・項目Aの a, b に共通する項目Bの501~1000は上位項目Xで共用する。
・項目A の b, c に共通する項目Bの1001~1500は上位項目Yで共用する。
・上位項目の定義は別に記録する。

**共用化よるデータ量の大幅削減**

合計2004レコード
（データ削減率:約33％）

実際には
・項目Aと項目Bとの関係を記録するテーブル
・上位項目と項目Bとの関係を記録するテーブル
に分割して管理する。

# 提案技術のデータ圧縮率



$$\text{データ圧縮率}\quad D = 1 - \frac{\sum_{i=1}^{n}(M_i+N_i)+S}{\sum_{i=1}^{n} M_i \cdot N_i+S} = \frac{1-X}{1+Y}$$

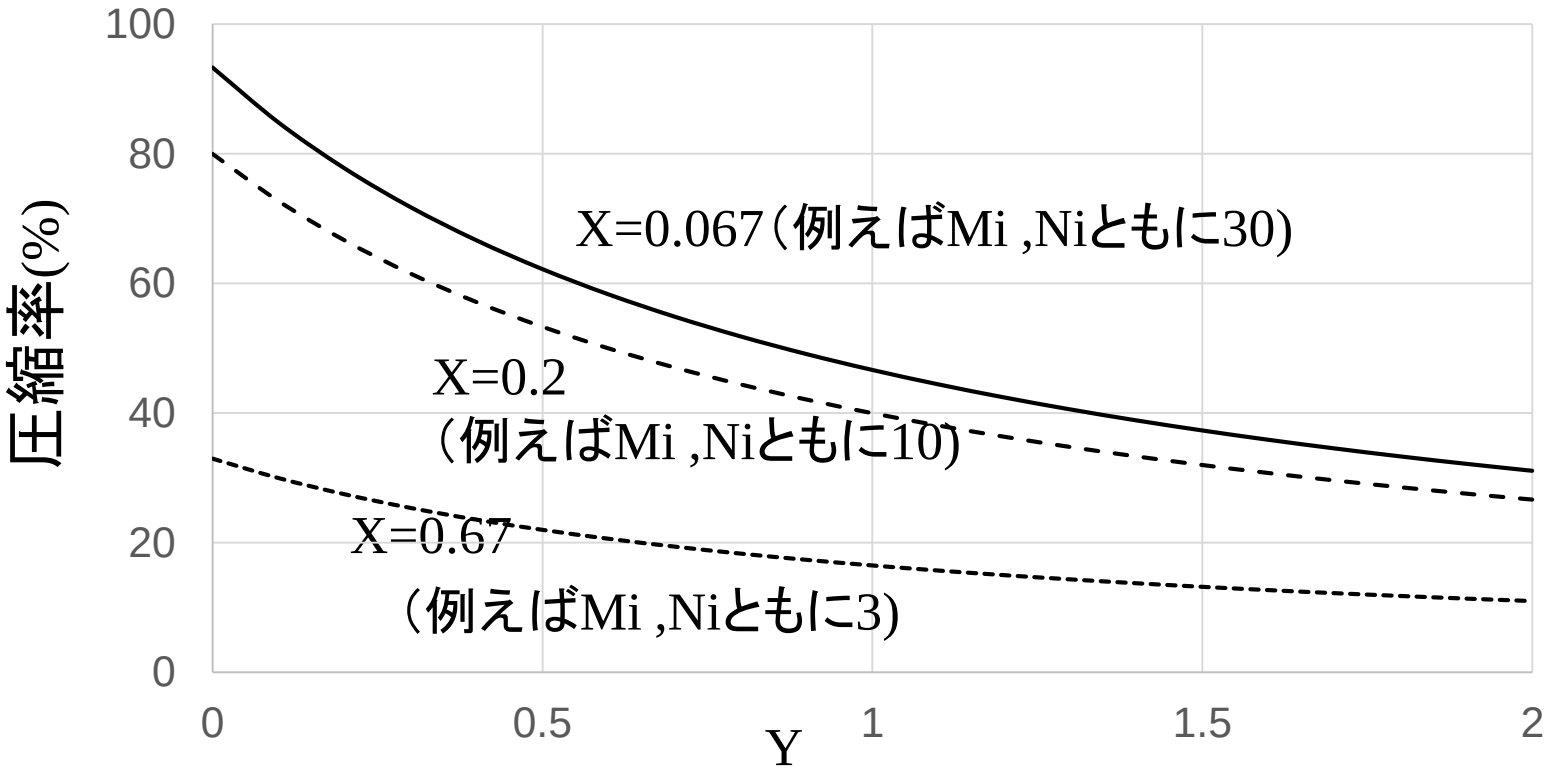
$$X = \frac{\sum_{i=1}^{n}(M_i+N_i)}{\sum_{i=1}^{n} M_i \cdot N_i}$$

$$Y = \frac{S}{\sum_{i=1}^{n} M_i \cdot N_i}$$

i: 上位項目番号
n:上位項目数
$M_i$:上位項目iに含まれる項目Aの件数
$N_i$:上位項目iに含まれる項目Bの件数
S: 上位項目で表現できない項目Aと項目Bの関係の数

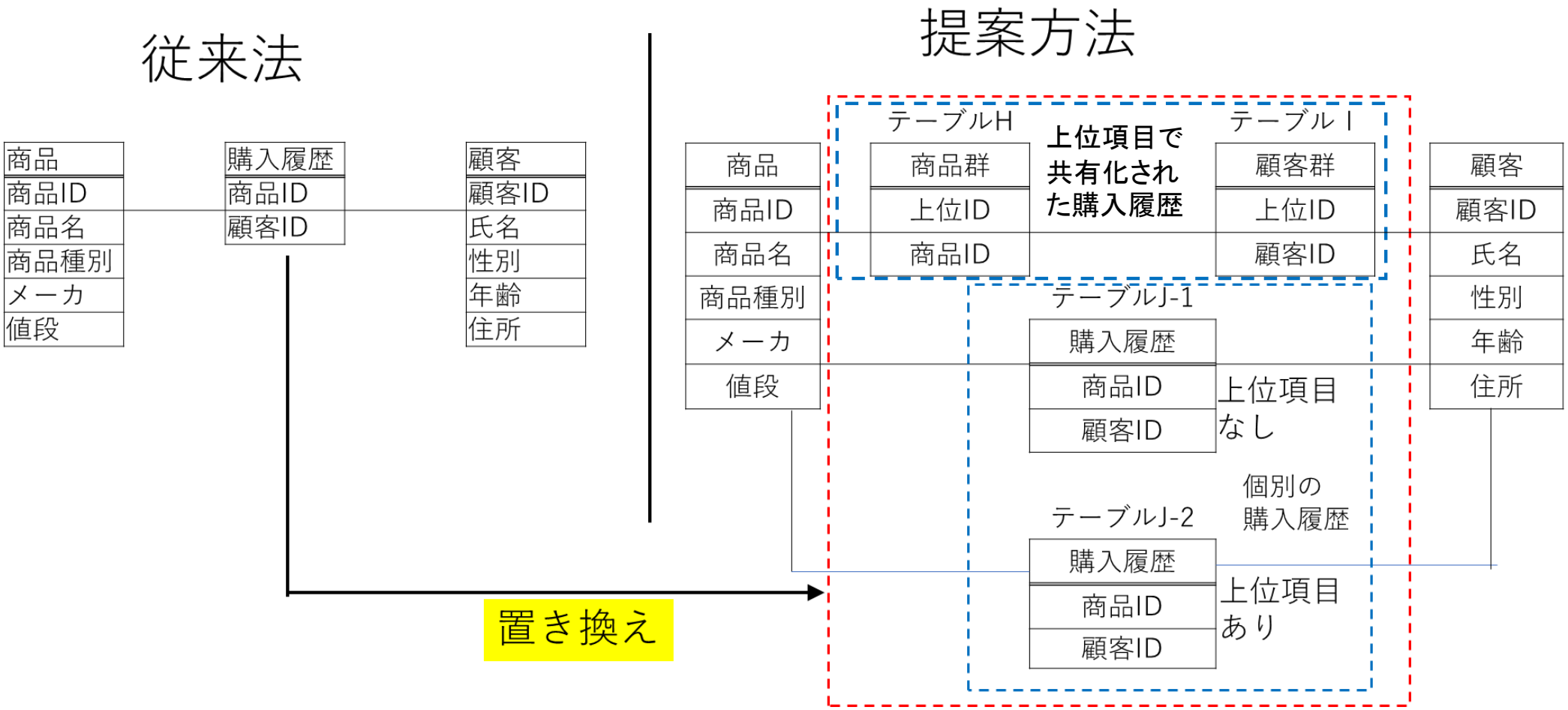
上位項目iに含まれる項目に限れば、従来方法のデータ量は$M_i \cdot N_i$という掛け算、
本方法のデータ量は$M_i + N_i$という足し算になる。

# データ圧縮率の計算例


圧縮率(%)
100
80
60
40
20
0
0
0.5
1
1.5
2
Y
X=0.067（例えばMi ,Niともに30)
X=0.2
（例えばMi ,Niともに10)
X=0.67
（例えばMi ,Niともに3)

# 具体的応用を想定したテンプレート

インターネットショッピングにおける購入履歴データを想定




例えば商品テーブルに1,000レコード、顧客テーブルに1,000,000レコードあれば、購入履歴テーブルには最大1,000,000,000レコードが存在する可能性がある。本方法は、そのように多数のレコード数が想定される多対多の関係を記録するテーブルに適用する。

# ケーススタディによる効果確認試験

新技術説明会
New Technology Presentation Meetings!

## 実験用データ

### 従来方法でのデータ

| 項目A | 項目B |
|---|---|
| 1~500 | 1~5,000 |
| 501~1,000 | 5001~10,000 |

合計5,000,000レコード

2個のインデックス付与
（項目A, 項目Bで個別にインデックス付与）

### 本方法でのデータ

テーブルH

| 上位項目 | 項目A |
|---|---|
| 1 | 1~500 |
| 2 | 501~1,000 |

テーブルI

| 上位項目 | 項目B |
|---|---|
| 1 | 1~5,000 |
| 2 | 5,001~10,000 |

合計11,00レコード

テーブルHのデータは
最初から最後まで不変

テーブルIはテータ入力
進展につれて順次増加

テーブルJ-1,J-2群は
入力終了段階では空。
途中段階ではデータが
存在する時あり

# ケーススタディによる効果確認試験

## 実験環境

- 従来構造にはインデックス付与
（2つのカラムの各々に個別にインデックス付与）

- 提案方法でもインデックス付与

- 使用RDB:MySQL5.1

- JAVAとJDBCを用いてRD操作するプログラムで実験
（OSはRedHat7)

# ケーススタディによる効果確認試験

新技術説明会
New Technology Presentation Meetings!

## 実験結果

| | | 従来法 | 本方法 | 従来法との比較 |
|---|---|---|---|---|
| データ量（レコード数） | | 5,000,000 | 11,000 | ◎ |
| 検索時間（秒/1000件） | ItemAで検索（出力5000件/1検索） | 2.5 | 1.3 | ○ |
| | ItemBで検索（出力500件/1検索） | 0.35 | 0.12 | |
| データ入力（秒/1000件） | | 0.03 | 0.10 | × |

# 想定される用途

(1) 個別アプリケーションでのデータ構造設計と
データ検索、更新（入力）用プログラム開発

特に多対多の関係で大量の記録が生じる場合に有効
（商品と購買者、Webサイトと訪問者、など）

(2)本方法によるデータ検索、更新機能を組み込んだ
汎用DBMSの開発

→ 個別アプリケーション開発の容易化
→ 本提案の長所の一層の強化

(3) 検索エンジンへの組み込み

検索記事とそのインデックスの関係の記録に適用

# 実用化に向けた課題

技術

・本方法の有効な適応範囲の確認

実用に近い複数のケーススタディでの評価

・既存フレームワーク（DBMS、検索エンジン）への組み込み

マーケッテイング

・本方法で効果の大きな具体的応用例の探索

開発体制

・パートナー企業の探索

# 企業への期待

(1) 本方法の効果的な応用事例*の提案と、
そのためのアプリケーション開発、ないし共同開発

(2) 本方法を組み込んだDMBSの開発、ないし共同開発

(3) 検索エンジンへ適用のための開発、ないし共同開発

(4) 上記(1), (2),(3)における開発体制の構築**

- データの更新頻度が検索頻度に比べて少ないか、検索の行われない時間にまとめて更新できるような用途、既に確定したデータの保存（例えば年度単位の記録のアーカイブ）などに特に好適

** (1)ではユーザ企業＋ITベンダ＋研究機関（大学等）
(2) ではDBMS開発会社＋研究機関（大学等）
(3) では検索エンジン運営会社＋研究機関（大学等）

# 本技術に関する知的財産権

(1)日本国登録　特許第6269884号
　2017年5月出願、2018年1月12日登録

(2)日本国出願　特願2018-090308
　2018年5月出願、公開前

(3)国際優先権出願　PCT/JP2018/018419
　2018年5月出願、2018年11月公開予定

(4)日本国出願　特願2018-16675
　2018年9月６日出願、公開前

いずれも出願人は神奈川大学、発明者は桑原恒夫

# 本技術に関する学会発表等

新技術説明会
New Technology Presentation Meetings!

(1)Kuwabara, T.: "New Data Structures to Reduce Searching Time on Databases“, IEICE General Conference 2018, D-4-7, p28 (2018)

(2) Kuwabara, T.: “New data structures to reduce data size and search time”, FIT2018　1D-1, No2, pp1-4 (2018)

(3) https://www.sci.kanagawa-u.ac.jp/info/kuwabara/PDF/FIT2018ForHP.pdf
(2) の内容の、学会の許可を得ての桑原研究室HP内への掲載

# お問い合わせ先

神奈川大学 研究支援部 平塚研究支援課付
調査役　栗林　健二

ＴＥＬ 0463－59－4111（代）
e-mail: sankangaku-renkei@kanagawa-u.ac.jp