情報統合と知識発見による高度情報利用

2004年6月11日

筑波大学
計算科学研究センター計算情報学部門
（システム情報工学研究科）
北川 博之
Email: kitagawa@cs.tsukuba.ac.jp

概要

- 大規模データ管理と
  データベース研究の流れ
- 情報統合と知識発見
- 現在の研究の一端を紹介
- まとめ

情報技術を取り巻く環境

- 情報機器の高性能化，大容量化，低価格化
- インターネットによる広域分散環境の実現
- マルチメディアの一般化
- モバイル・ユビキタス環境の進展
- 情報の処理／通信／放送の融合

⇒ あらゆる人間の活動が情報技術とは無関係には存在し得ない時代

デジタルデータの急増

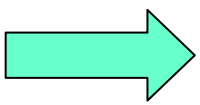
- “How Much Information? 2003”
  - カリフォルニア大学バークレー校
    P. Lyman & H. R. Varian
  - 2002年に新規に生み出された情報の量
    - $5x10^{18}$バイト=5エクサバイト
    - 米国会図書館の蔵書の情報量の約50万倍
  - 92%の情報は磁気的メディア（大部分はディスク）に格納されたもの

デジタルデータの急増

- Storage Law
  - 世界中のデジタルストレージの総容量は，9ヶ月で倍増
  - Mooreの法則よりも急激な増加
- 種々のバズワード
  - Data Tombs
  - Write-only Data
  - Data Tsunami

計算科学における大規模データの重要性

- 観測・実験データ
- 数値シミュレーションデータ
- 大規模かつ多様なアーカイブ，データベース
- 科学技術文献情報・特許情報
- メタデータ・オントロジー・タキソノミー
- シミュレーションモデル，プログラム
- 解析ツール群

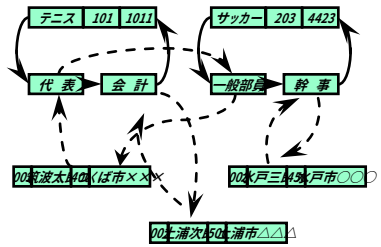


「仮説形成，理論形成，実験，検証」の全てにおいて大規模データやオブジェクトの統合的利用が重要

データベース研究の流れ

- 1960年代
  - ネットワーク型／階層型DBMS
- 1970年代
  - リレーショナルデータモデルの提案（1970 E. F. Codd）
  - リレーショナルDBMSの実現技術
  - データモデル論，データベース設計論
  - ACM SIGMOD, VLDB
- 1980年代
  - リレーショナルDBMSの実用化
  - ポストリレーショナルデータベース研究

    オブジェクト指向，分散DB，並列処理，知識処理, …



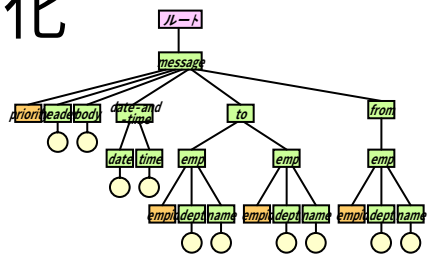
サークル名	部屋番号	部屋内線番号
テニス	101	1011
サッカー	203	4423

社員番号	氏名	基本給与	住所
001	筑波太郎	400	つくば市×××△
002	土浦次郎	500	土浦市△△△
003	水戸三郎	450	水戸市○○○

科目番号	科目名	単位数	担当者	実習課題	
			担当者	課題番号	課題名
001	データベース	2	北山 山田	01 02 03	データモデリング データベース設計 SQL
002	システムプログラム	3	鈴木 佐藤	01 02	Cプログラミング システムコール

データベース研究の流れ

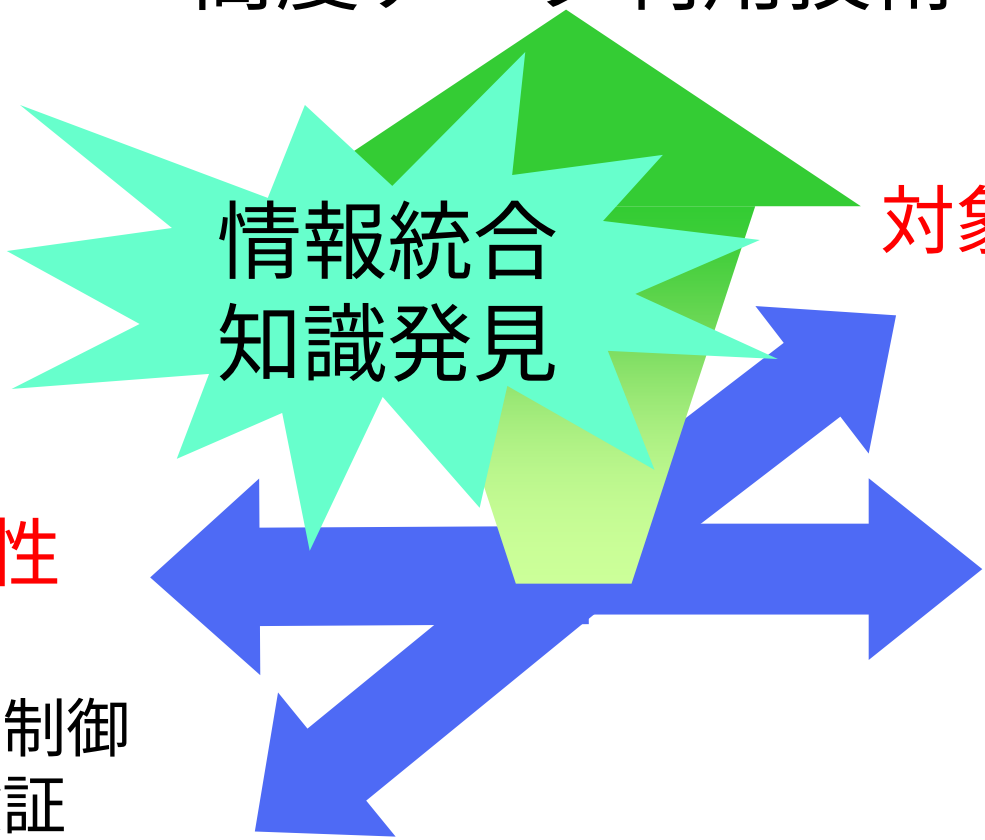
- 1990年代
  - オブジェクトリレーショナルDBの実用化
  - インターネット，WWWの普及
  - XML，半構造データ
  - 情報検索の復権とデータベース技術との融合
  - モバイル・ユビキタス環境におけるデータ管理
  - データウェアハウス，データマイニング



- トップダウン的アプローチからボトムアップ的アプローチへ
  - グローバルかつオープン環境におけるデータ利用
  - 大量データからの情報獲得を支援する技術

データベース研究の展開

高度データ利用技術



対象データ

メタデータ
マルチメディア
XML・Web

処理機能

トランザクション
類似検索
コンテンツ分析

信頼性

リカバリ
同時実行制御
整合性検証

性能向上

問合せ最適化
並列処理

情報統合

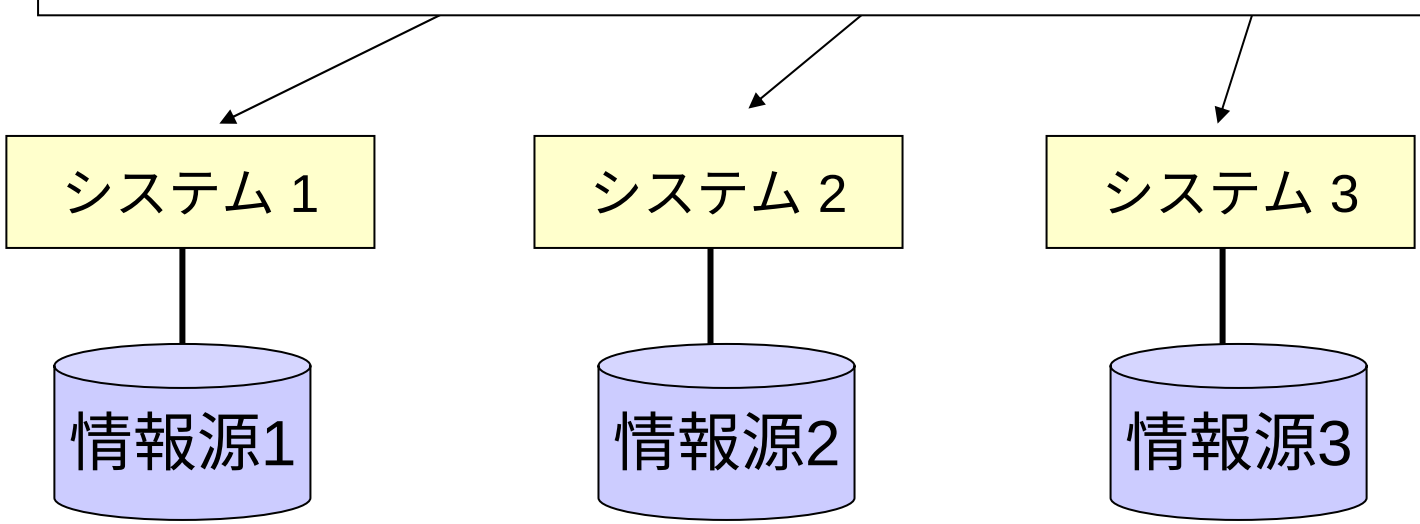
- 背景
  - 情報統合はデータベース出現の元々の要因
    →「データベース研究にとっては永遠の課題」
  - ネットワーク環境の進展に伴う分散環境
    - 分散データベース，マルチデータベース
  - 多様な情報源の統合利用
    - RDB，テキスト，Web，マルチメディア
- アプローチ
  - メディエータ／ラッパー
  - データウェアハウス

情報統合の必要性

SQL, XQuery,
Google API, …

データアクセス法の違い
データ形式の違い
メタデータの記述や所在の違い
情報源探索の必要
異なる情報源中のデータを関連づける方法の欠如
等の種々の問題

利用者



メディエータ／ラッパー

利用者

メディエータ

統合データモデル

ラッパー 1

ラッパー 2

ラッパー 3

システム 1

システム 2

システム 3

情報源1

情報源2

情報源3

データウェアハウス

利用者

データウェアハウス

データウェアハウスサーバ

エクストラクタ 1

エクストラクタ 2

エクストラクタ 3

システム 1

システム 2

システム 3

情報源1

情報源2

情報源3

2つのアプローチの比較

	メディエータ／ラッパー	データウェアハウス
情報源アクセス	要求駆動型	事前抽出型
データの鮮度	最新	抽出時
ローカル処理への影響	大	小
グローバル処理性能の保証	難	可能
その他	情報源の変更，動的統合への対応がしやすい	• 履歴情報の蓄積が可能 • データウェアハウス管理が必要

統合化された情報の利用

- 集約的データ処理
  - 問合せ／集計計算／レポート出力
- OLAP (On-Line Analytical Processing)

  cf. OLTP (On-Line Transaction Processing)
- データマイニング，知識発見

知識発見とデータマイニング

- 知識発見（Knowledge Discovery in Databases）:
  - 有効性，新規性，（潜在的な）有用性をもち，かつ人間が理解可能なパターンをデータから発見するプロセス
- データマイニング:
  - しかるべき水準の効率をもってデータから特定のパターンを抽出するために計算技術を適用する知識発見のプロセスの一部

[U. Fayyad: SSDBM97]

知識発見とデータマイニング

- データクリーニング
- データ統合（データウェアハウスへの格納）
- 分析対象データ選択
- 分析に適した形式へのデータ変換
- データマイニング→データパターンの抽出
- パターン評価
- 知識の提示

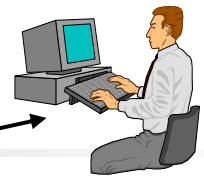
データマイニングの代表的手法

- 相関ルール (association rule)
  - データに内在する相関性のパターンを抽出
- 分類 (classification)
  - データを与えられたクラスのいずれかに分類
- 回帰 (regression)
  - 他の属性値からある属性値を予測
- クラスタリング (clustering)
  - データをその属性に基づき複数のクラスに分類
- 弁別 (discrimination)
  - あるクラスに属するデータの特徴を抽出
- 外れ値検出 (outlier detection)
  - 他のデータと性質が異なるデータを検出
- その他
  - テキストマイニング，Webマイニング，ストリームマイニング

当グループにおけるアプローチ

- 情報統合に関する研究
  - 異種分散情報源の統合
  - タキソノミーを用いたウェブサーチ技術
  - 情報統合のためのインタフェース

異種情報源統合



クライアント

(2) 統合スキーマ

(3) 統合スキーマに基づく問合せ

WebNR/SD
拡張リレーショナルモデルに
基づくメタデータベースシステム

(6) 最終問合せ結果

メディエータ

(1) 局所スキーマ

(4) 各情報源に対する問合せ

(5) 部分解

ラッパー

ラッパー

ラッパー

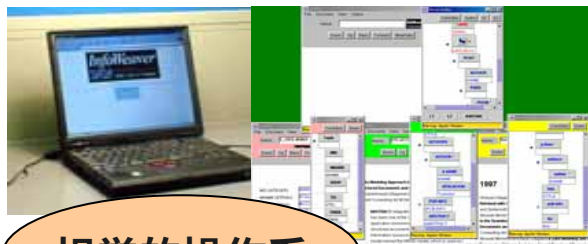
RDB

Document
Repository

Web

情報統合システムInfoWeaver

メディエータ



RMI

視覚的操作系

Oracle

ラッパー

ラッパー

ラッパー

文書検索システム
OpenText

Web

名前	TEL	EMail
佐藤	9512	Sato
山田	9643	Yama

佐藤,
”Integration of Web”,
Abstract

1996

1997

1998
佐藤,
”Integration of Web”,
Proc. ABC.
山田, “Data…”

DB研ホームページ

リンク

リレーショナルデータベース

フルテキストデータベース

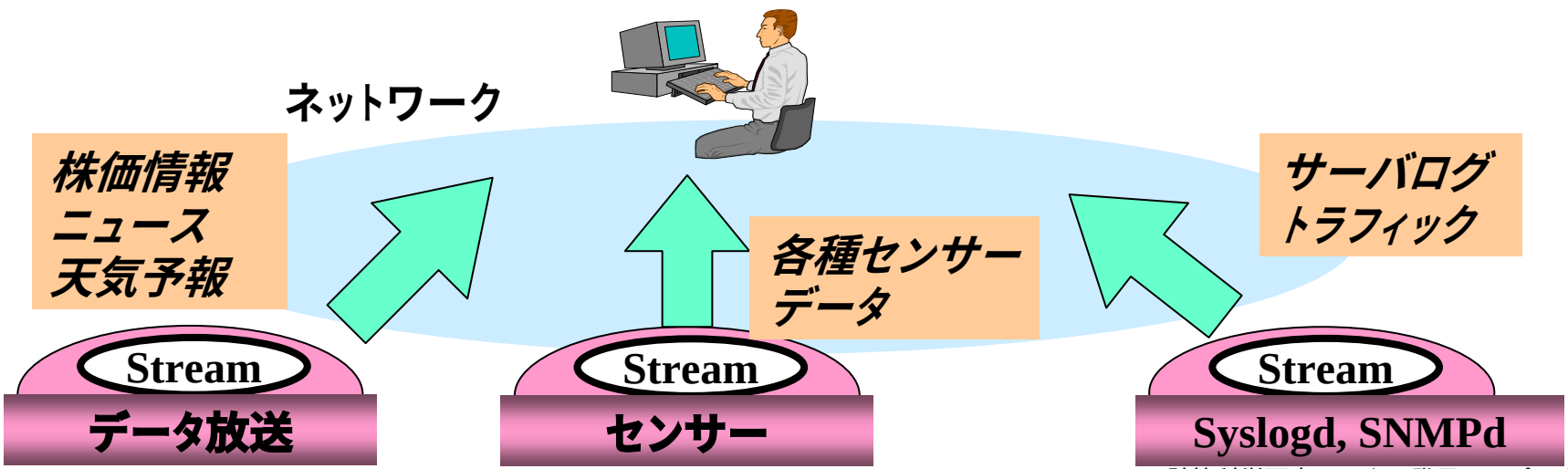
Webページ群

データストリームを含めた情報統合

科研費特定領域研究

- ネットワーク技術の発達
- センサー，計測デバイスの小型化・低価格化

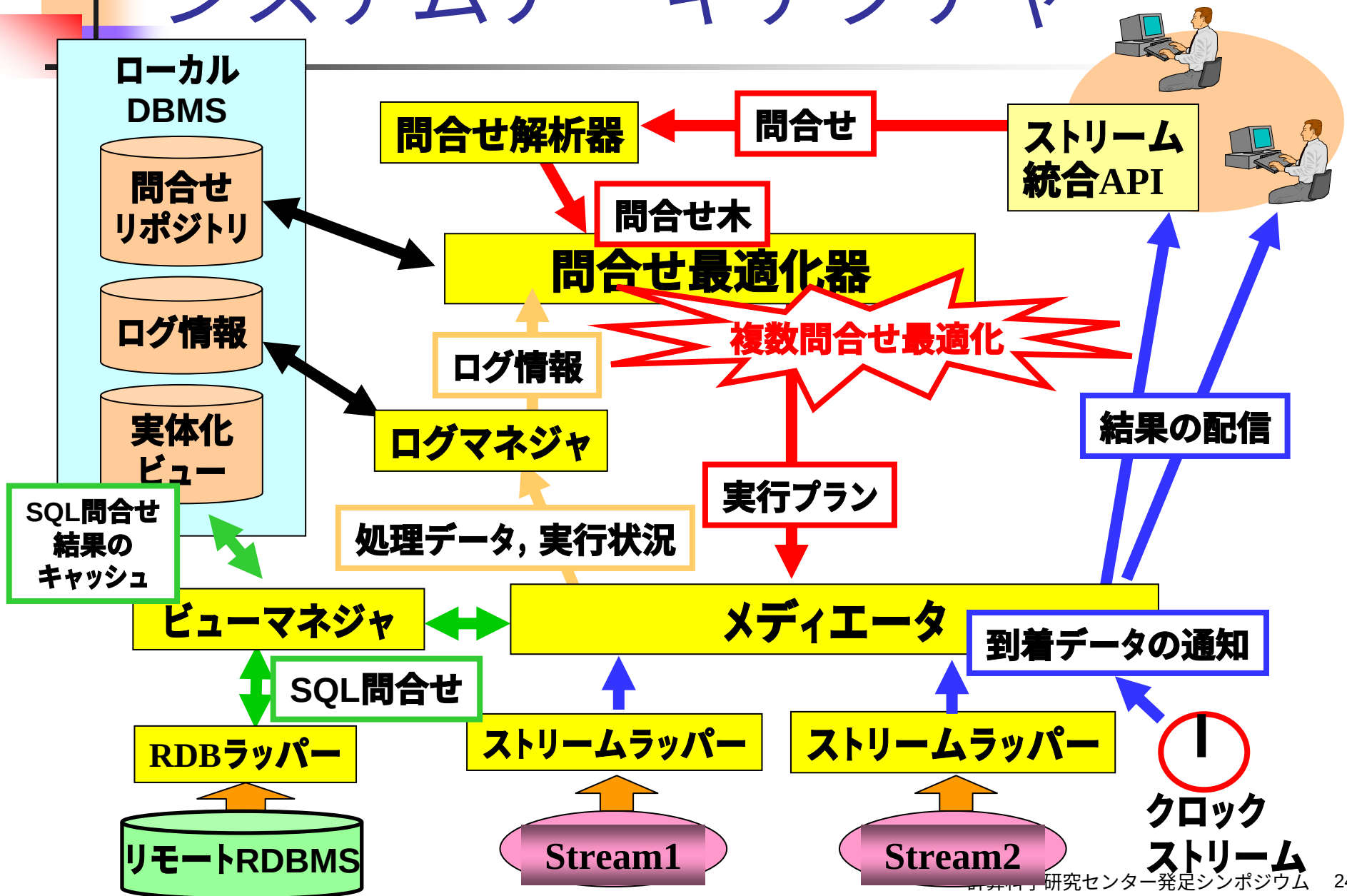
- 大量のデータストリームが利用可能
  - 時々刻々と変化する情報を逐次送ってくる情報源
  - センサーネットワーク，情報配信サービス，ログ情報
- データストリームの高度統合利用



利用例：観測情報提供システム

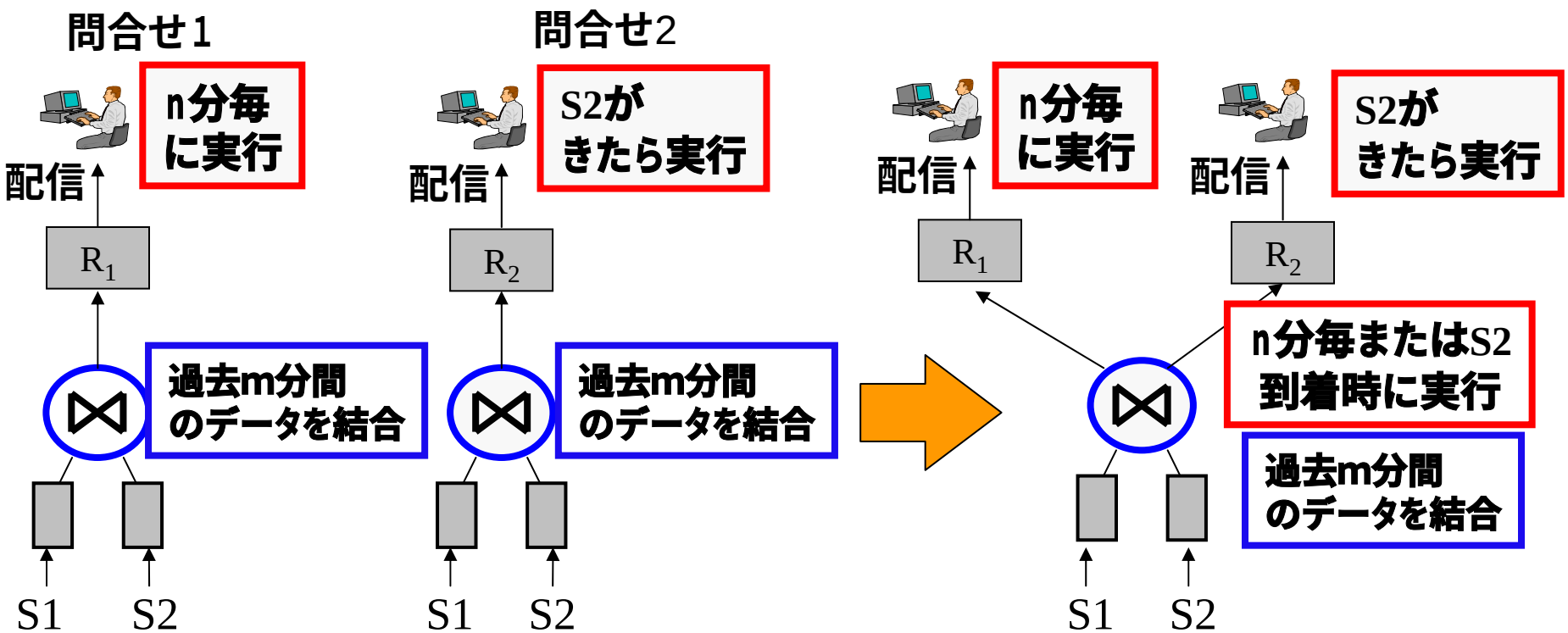
- 衛星からの観測データおよび地上の観測所のデータをリアルタイムに統合

システムアーキテクチャ



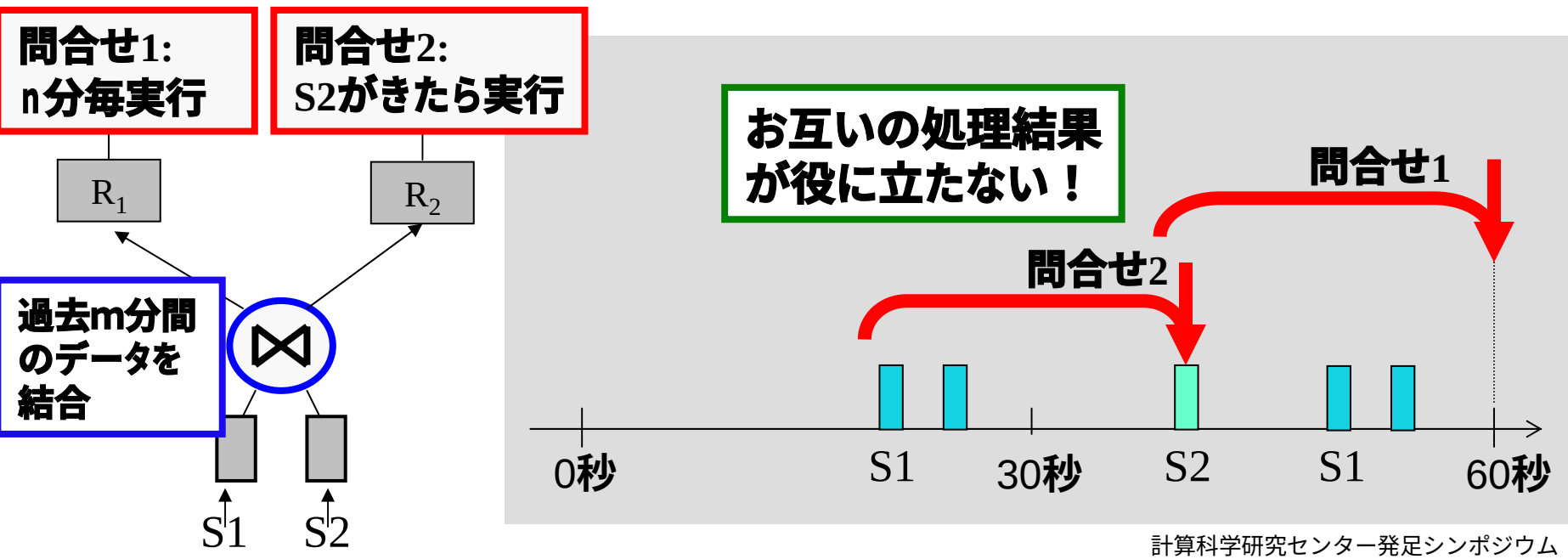
複数問合せ最適化

- 複数の問合せの中に含まれる共通演算に着目
  - 処理結果を共有することで効率化を図る



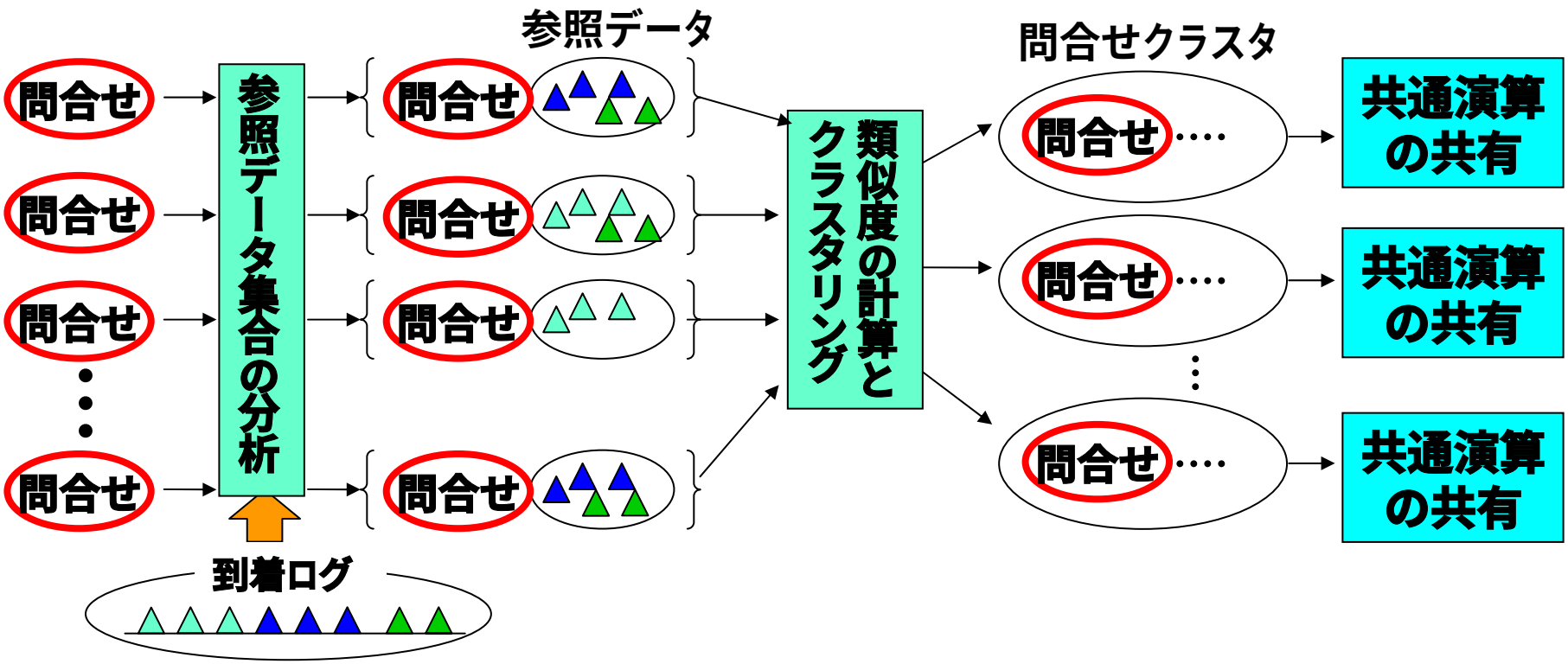
ストリームにおける複数問合せ最適化の注意点

- 実行タイミングが離れている場合
  - 異なる範囲のデータを参照してしまい，共有できるデータが生成されないかもしれない



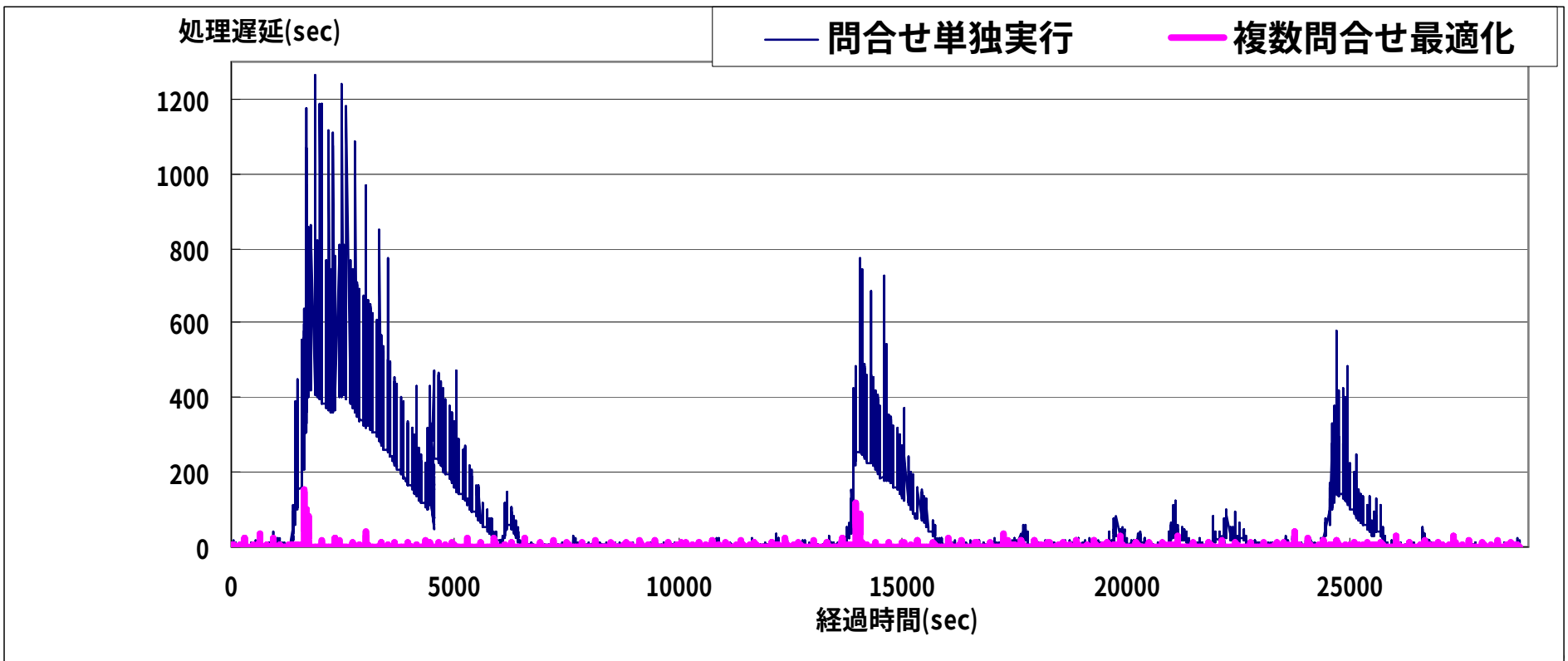
複数問合せ最適化

- ストリームデータの到着パターンをマイニングすることで問合せのクラスタを生成
- クラスタ内では中間結果を共有



予備的実験評価

- データが到着してから必要な処理が完了するまでの時間
- 比較
  - ５００個の問合せを単体で実行した場合
  - ５００個の問合せに複数問合せ最適化を適用した場合



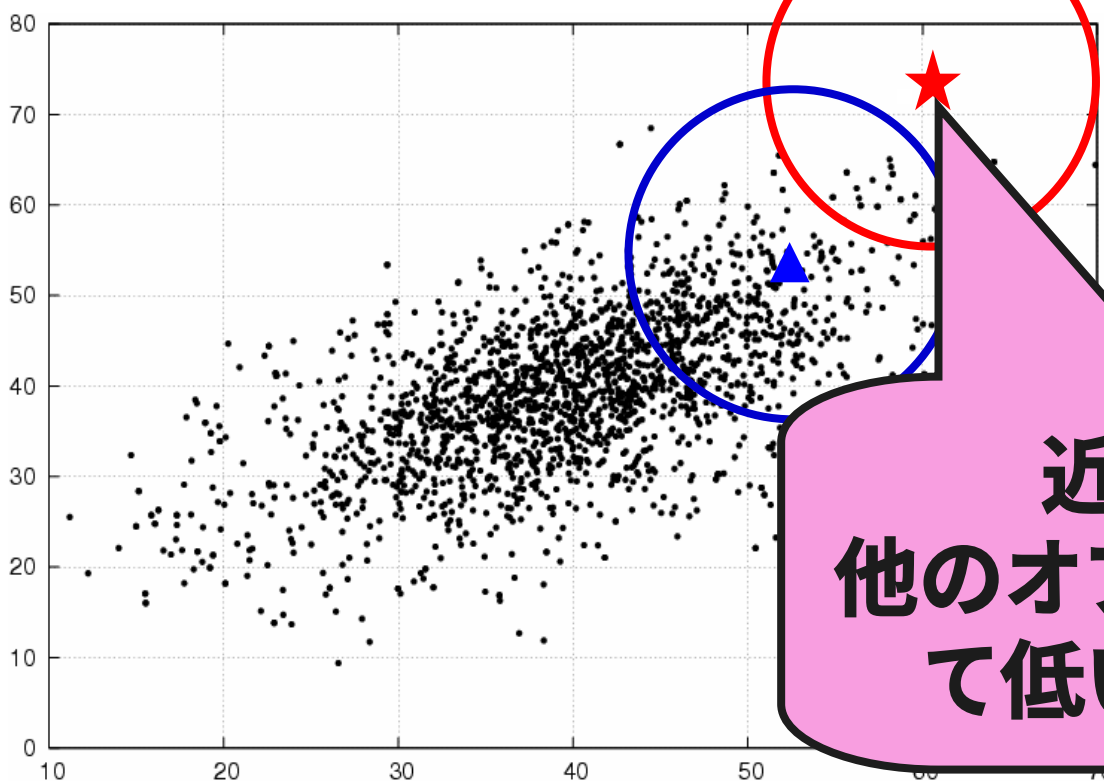
当グループにおけるアプローチ

- 情報統合に関する研究
  - 異種分散情報源の統合
  - タキソノミを用いたウェブサーチ技術
  - 情報統合のためのインタフェース
- データマイニング・知識発見
  - 利用者の意図を反映した外れ値検出
  - 空間情報源の発見のためのWebマイニング
  - テキストストリームからのトピック抽出

外れ値検出

科研費基盤研究，学振日米共同研究

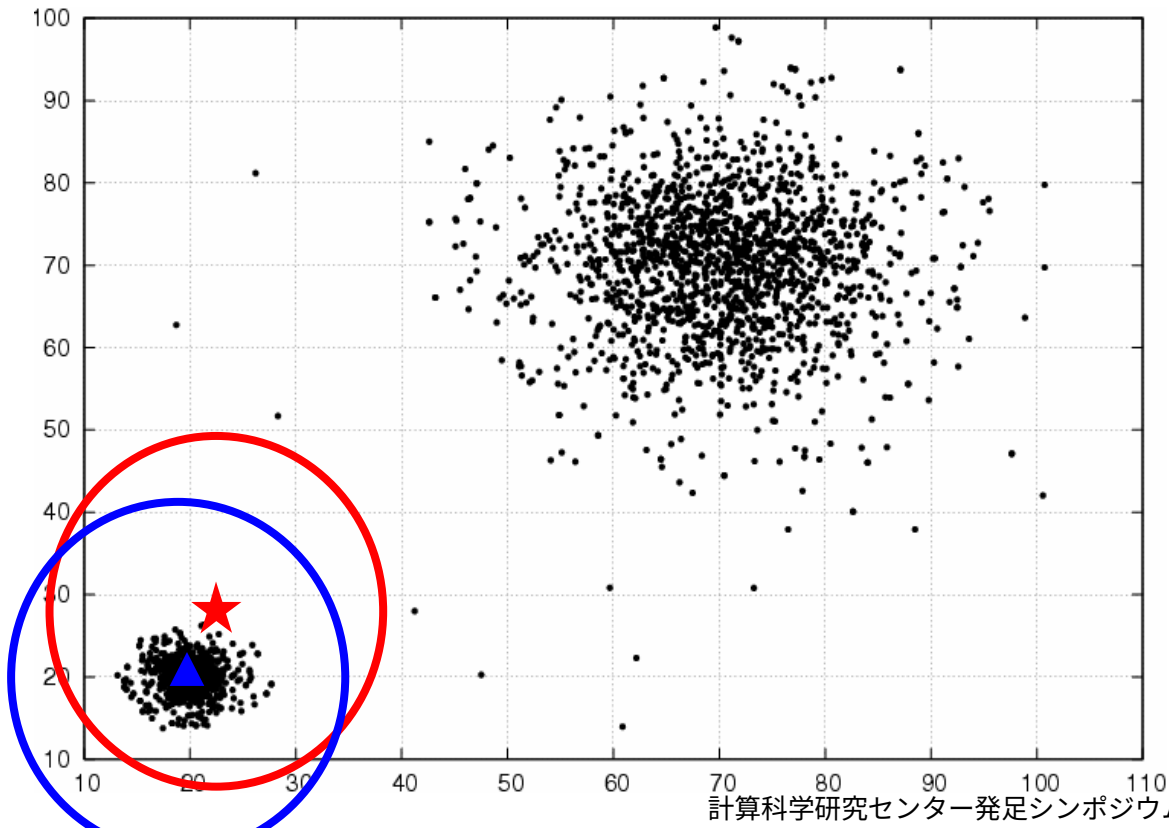
- **外れ値 (Outlier) : 他のオブジェクトに比べてその振る舞いが大きく異なるもの**



近傍密度が他のオブジェクトに比べて低いので外れ値

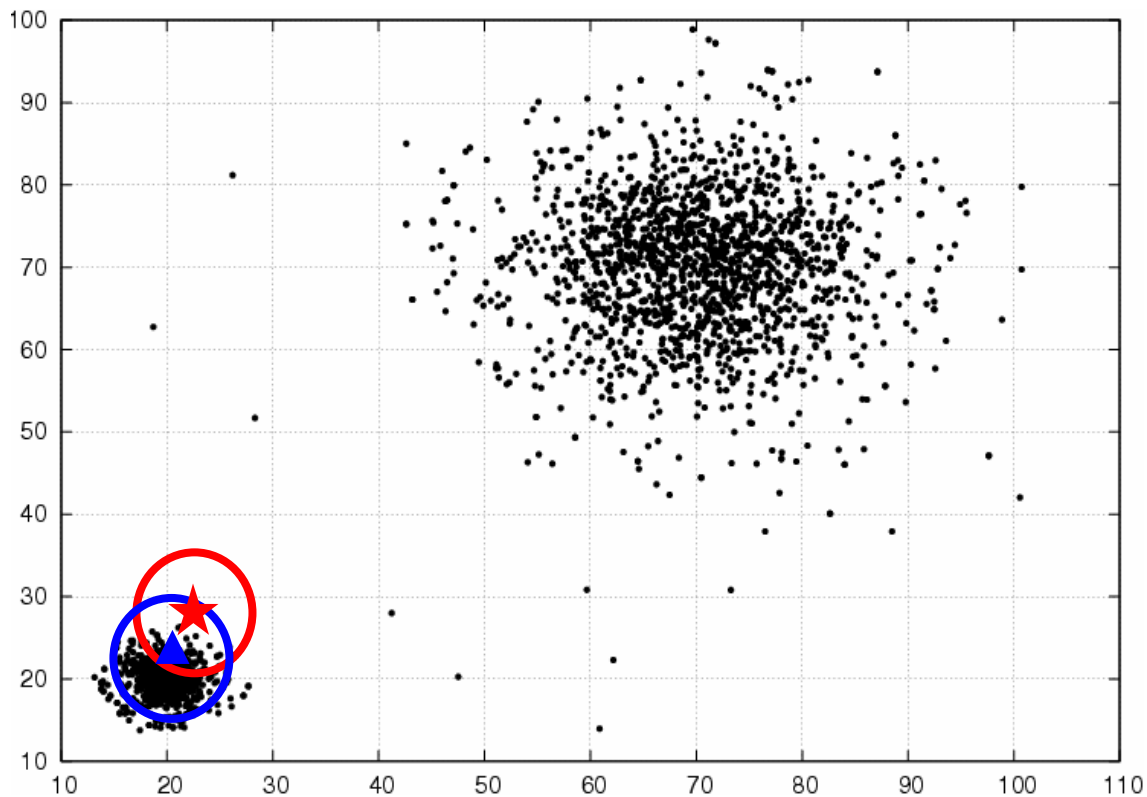
異なったスケールにおける外れ値

- **何を外れ値とみなすかは状況により変化**



異なったスケールにおける外れ値

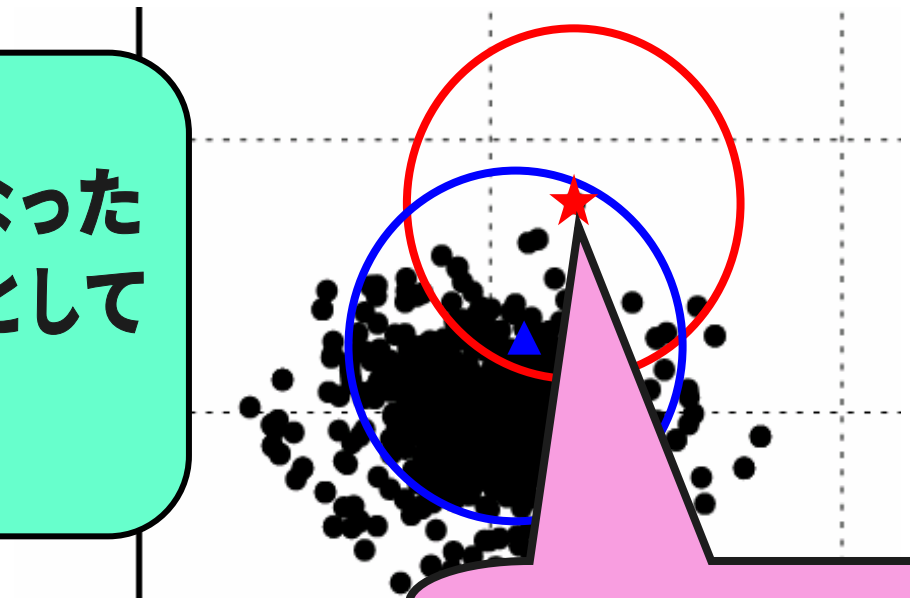
- 何を外れ値とみなすかは状況により変化



異なったスケールにおける外れ値

- 何を外れ値とみなすかは状況により変化

スケールに応じて異なったオブジェクトを外れ値として検出すべき

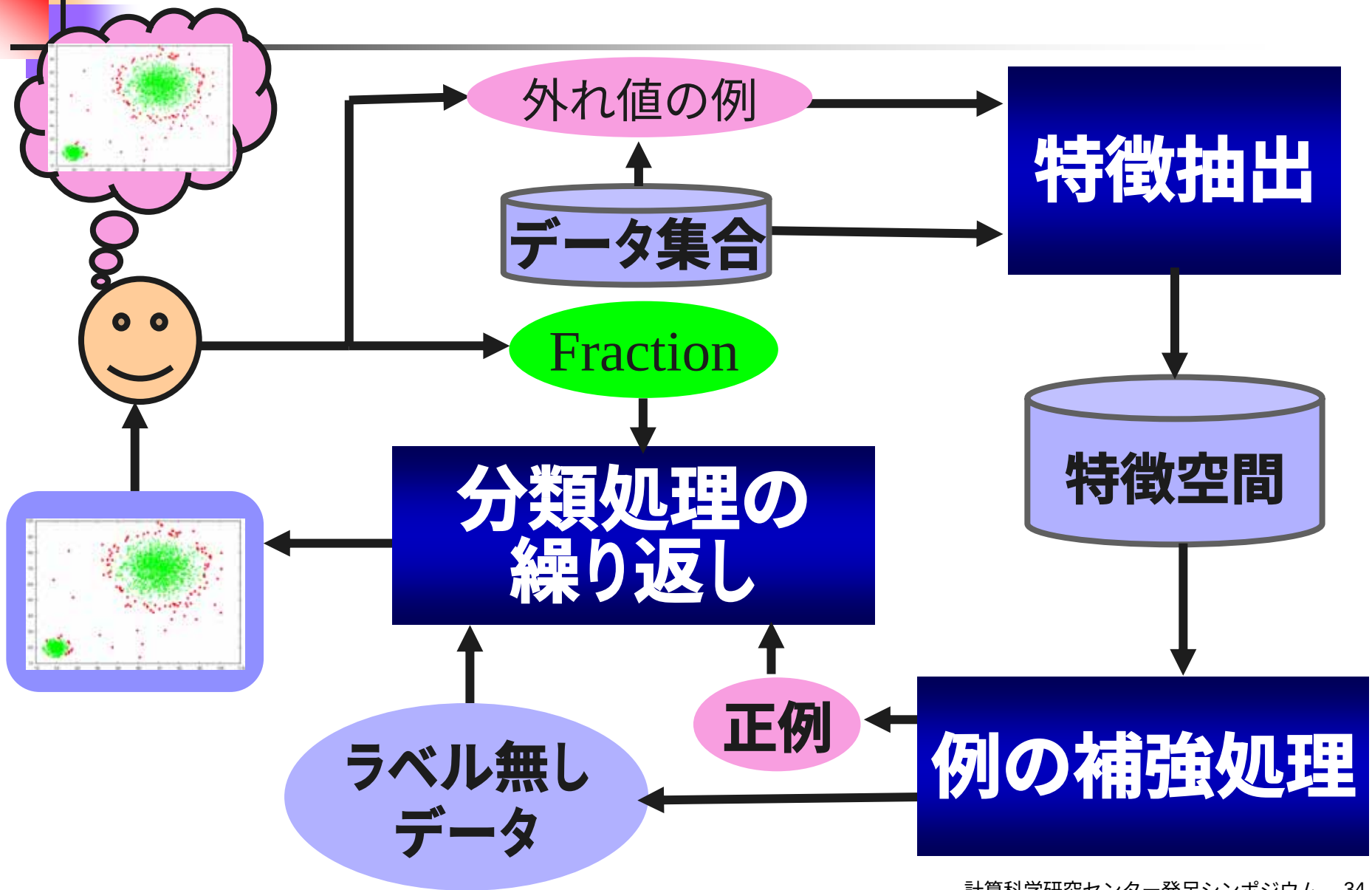


10

10

ミクロなスケールで見た場合は外れ値

例示に基づく外れ値検出[PAKDD04]

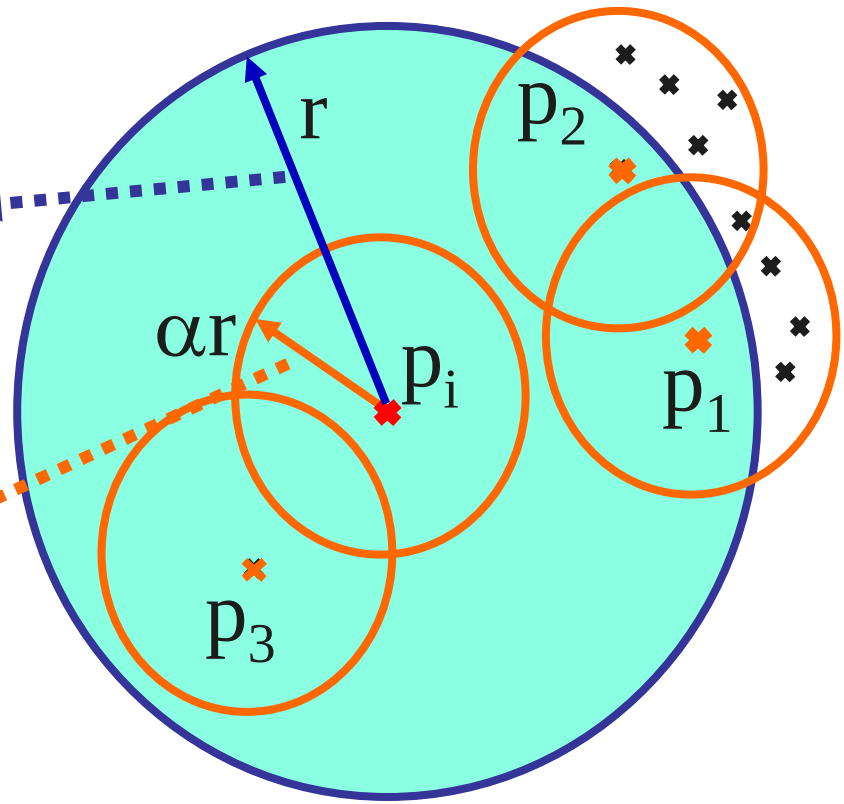


MDEF：外れ値とみなせる度合

$$MDEF(r, p_i) = \frac{\text{平均密度} - \text{近傍密度}}{\text{平均密度}}$$

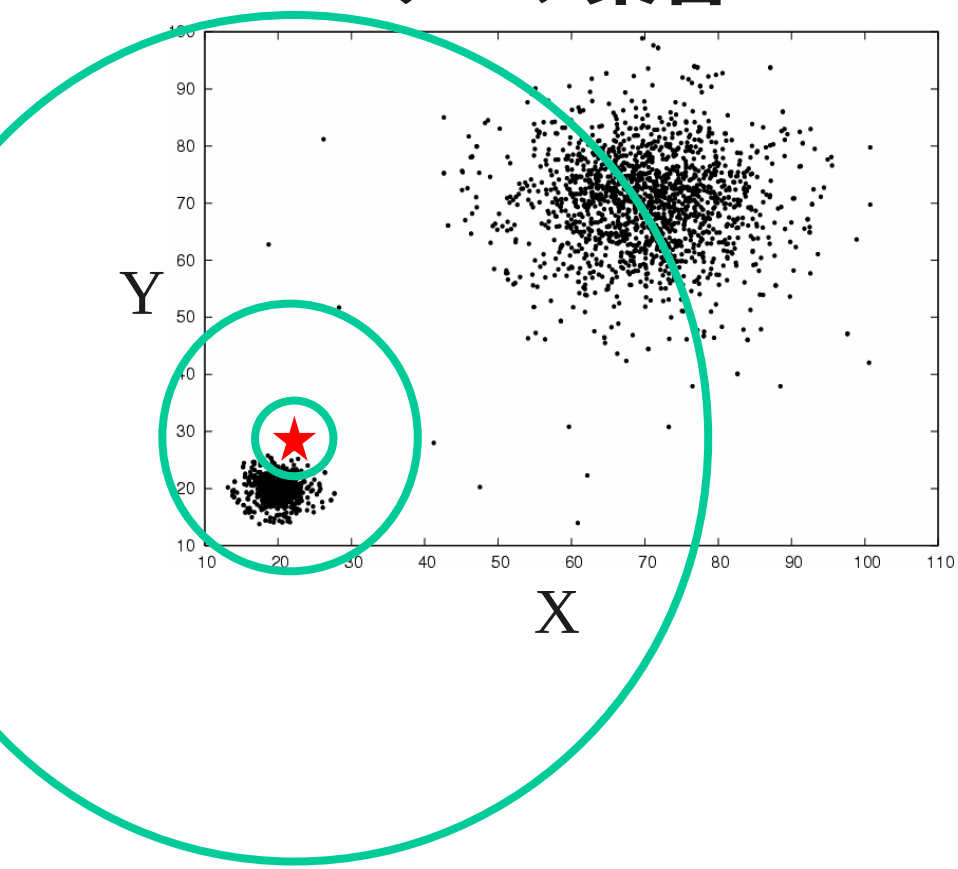
平均密度：
p_i の r-近傍内にあるオブジェクトの近傍密度の平均値

近傍密度：
p_iのαr-近傍内にあるオブジェクトの個数

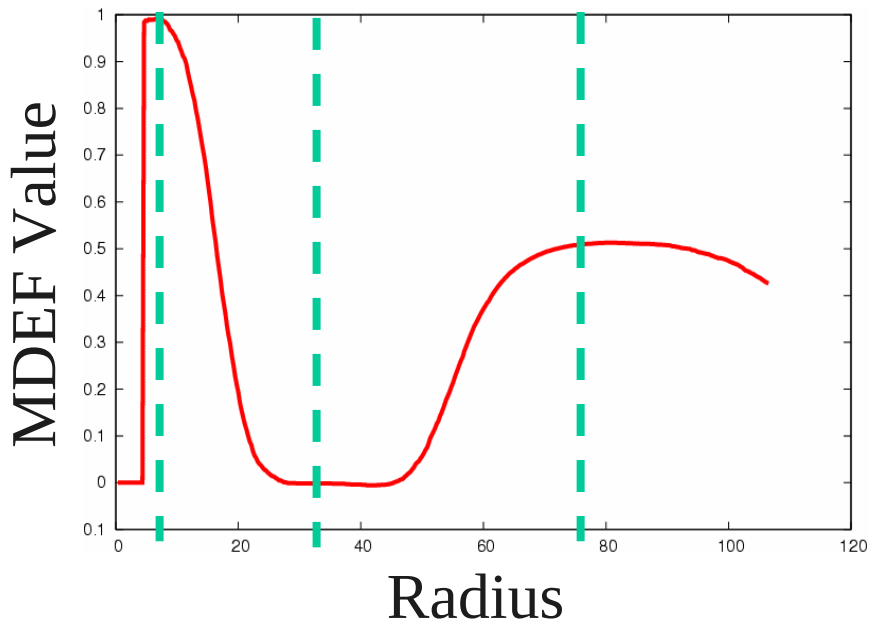


MDEFプロット

データ集合

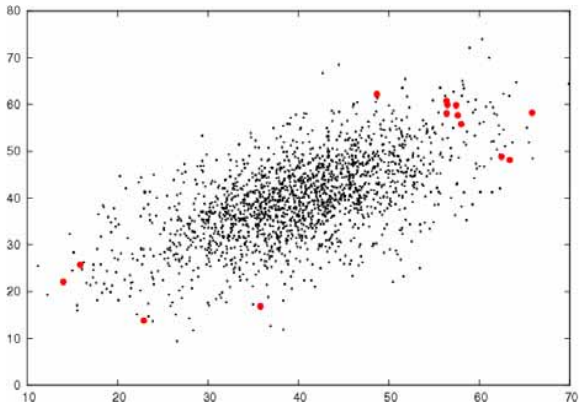


MDEFプロット

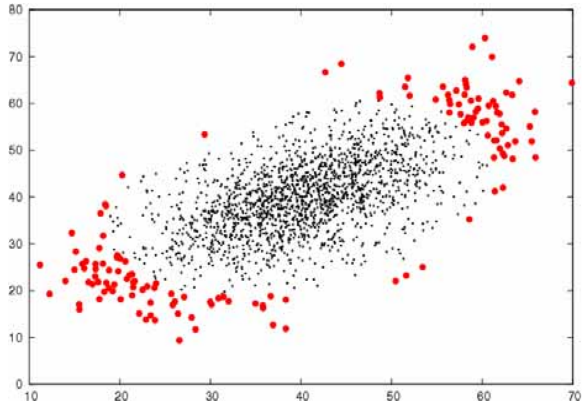


正規分布と外れ値

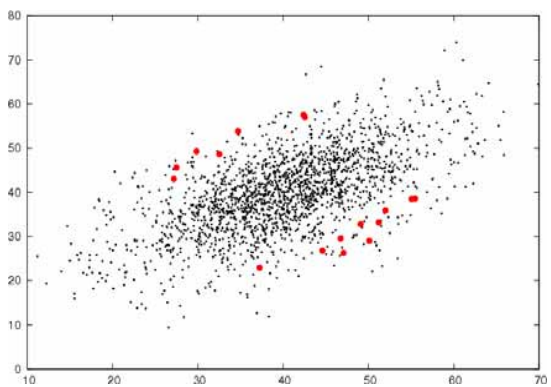
例示データ:

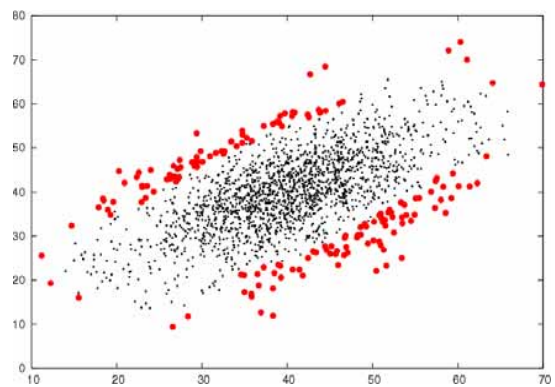


検出結果:



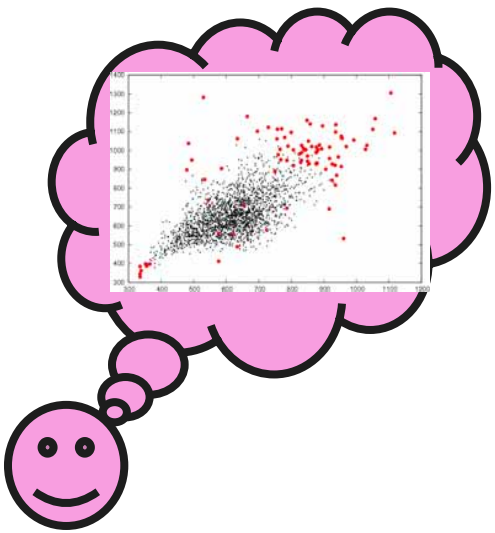
適合率=88.7%,**再現率**=92.1%



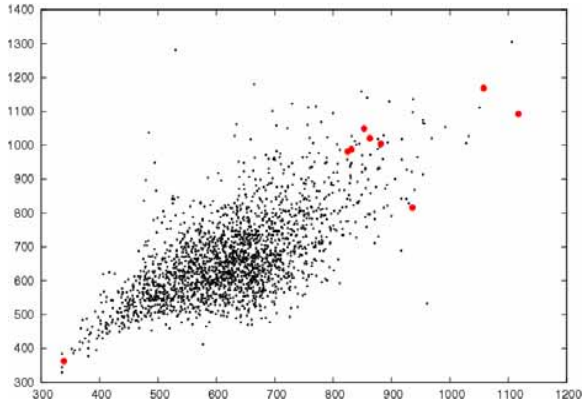


適合率=76.5%,**再現率**=80.0%

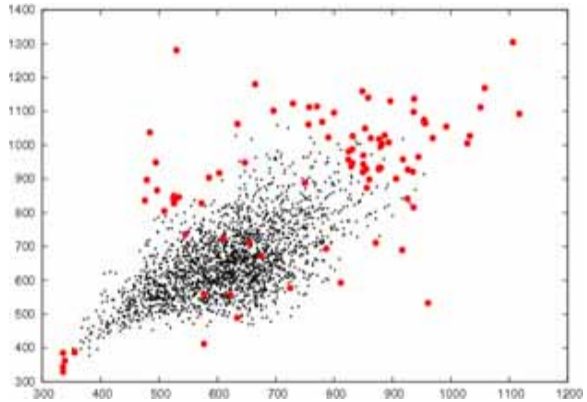
NY Women Marathon



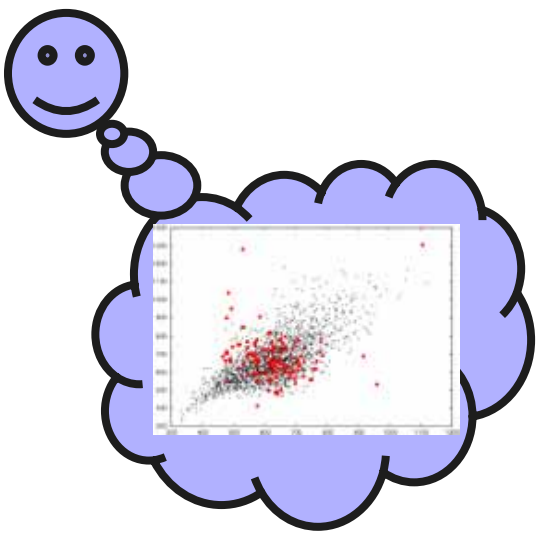
例示データ:

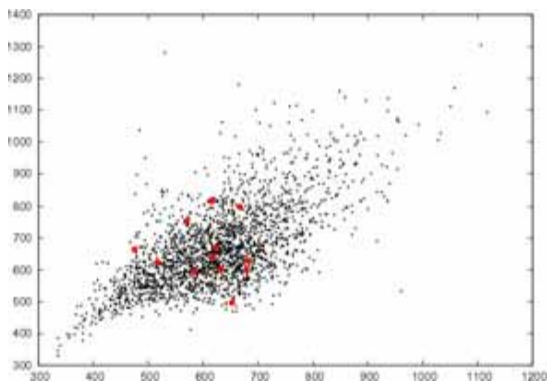


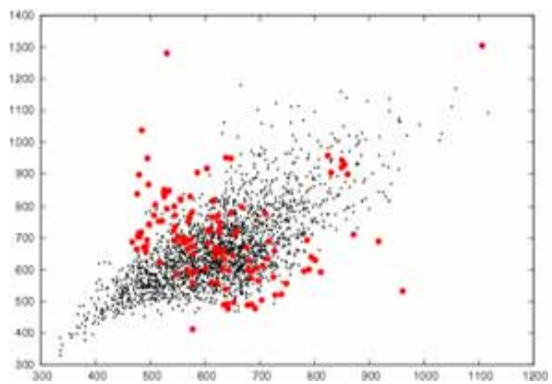
検出結果:



適合率=81.5%, **再現率**=85.0%

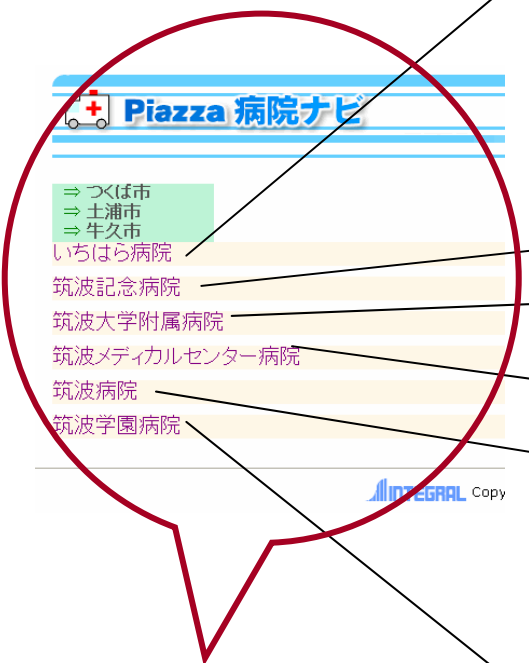




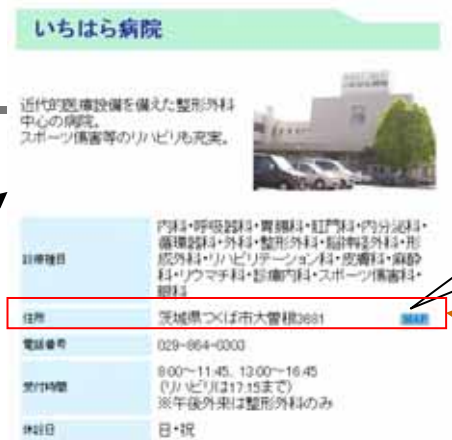


適合率=66.6%, **再現率**=70.7%

Web中の空間情報ハブ



空間情報ハブ



茨城県つくば市大曽根3681

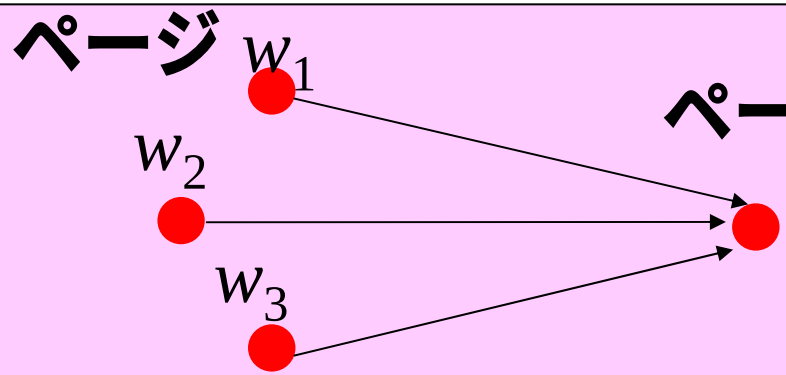
空間リンク

茨城県つくば市上横場2573-1

地理的空間

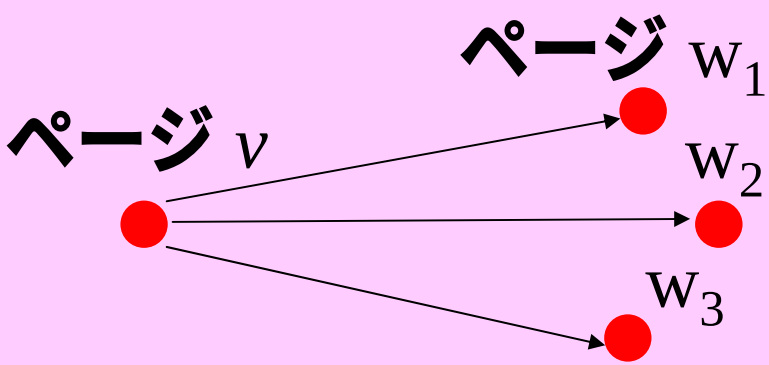
ＨＩＴＳ：ハブとオーソリティ

- 良いオーソリティページは多くの良いハブページに指されている



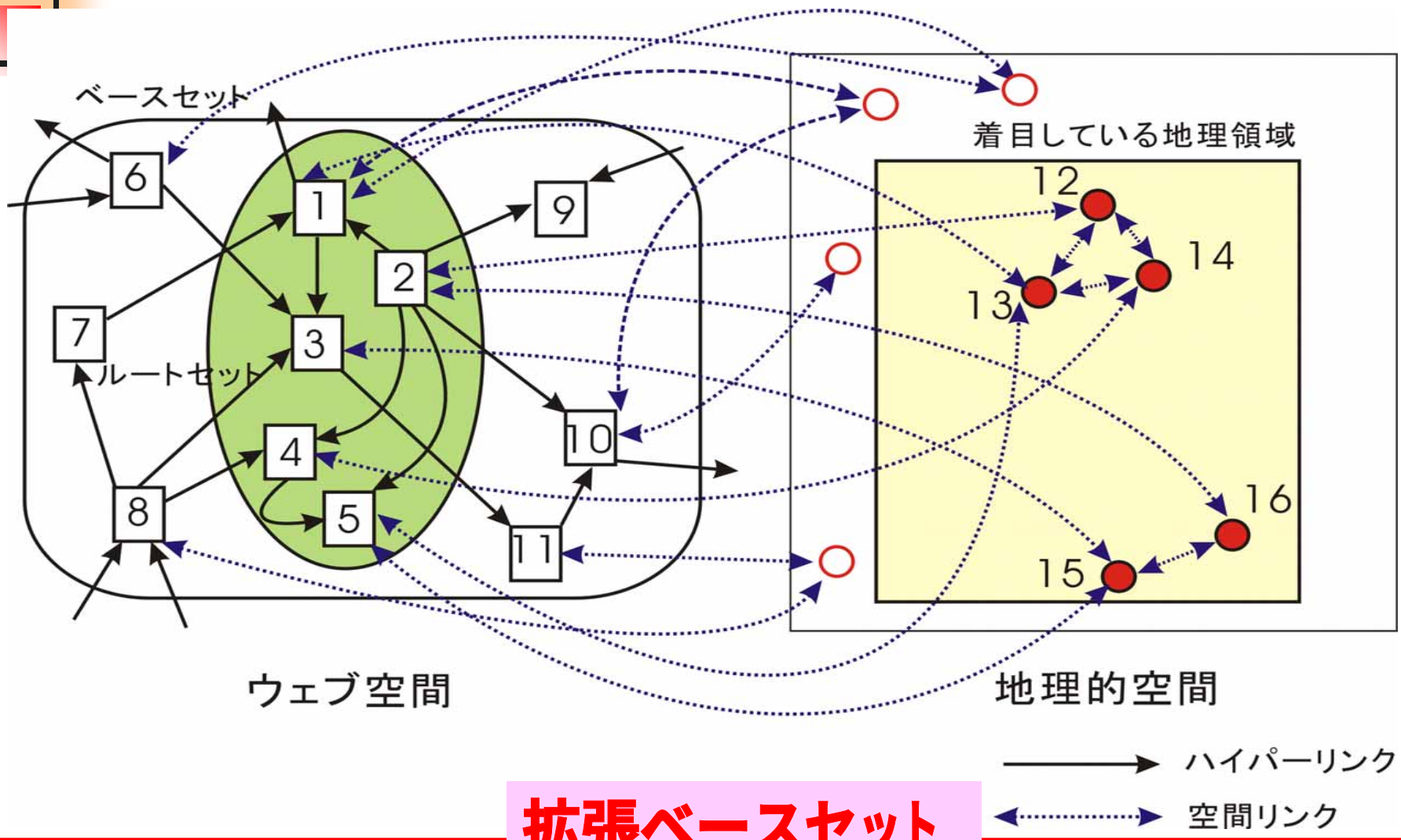
$$a(v) = \sum_{w \in pa[v]} h(w)$$

- 良いハブページは多くの良いオーソリティページを指している



$$h(v) = \sum_{w \in ch[v]} a(w)$$

拡張ベースセット



拡張ベースセット

ウェブページ、ウェブリンク、空間ノード、空間リンクからなる

予備実験

- NTCIR-4 WEBタスク文書データ
- 主として.jpドメインから2001年に収集したHTMLもしくはプレーンテキストファイル , 約1100万件,リンク数約8000万
- 空間情報の抽出
  - 郵便番号
  - 空間情報と経緯度の対応付け

当グループにおけるアプローチ

- 情報統合に関する研究
  - 異種分散情報源の統合
  - タキソノミを用いたウェブサーチ技術
  - 情報統合のためのインタフェース
- データマイニング・知識発見
  - 利用者の意図を反映した外れ値検出
  - 空間情報源の発見のためのWebマイニング
  - テキストストリームからのトピック抽出
- Webコンピューテイング
  - P2P環境における効率的情報検索
  - XMLデータベース，XMLデータ処理

まとめ

- 大規模データの高度利用
  - 情報統合：分散，異種インタフェース，異種メディアの統合利用
  - 知識発見：膨大なデータの効果的利用
- 今後の展開
  - 計算科学はこれら技術の実践と発展の場
  - 計算科学におけるデータ利用に関わる問題への適用と新たな研究課題の発見
  - 先端的大規模データ管理・利用技術の研究開発
  - 異分野研究者の連携

ご清聴ありがとうございました.