能動的リソースマイニングに基づく異種情報統合基盤の研究

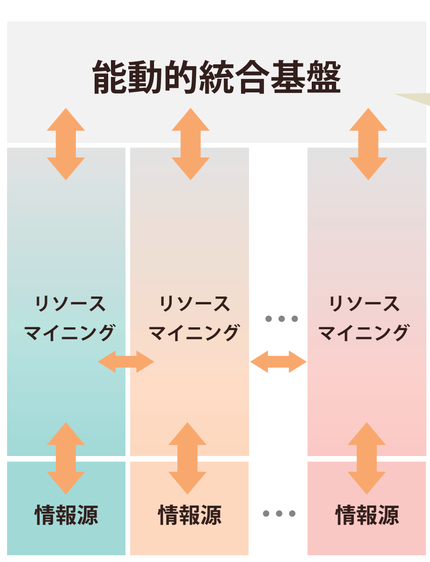
研究代表者：北川博之（筑波大学）　分担者：天笠俊之，森嶋厚行，古瀬一隆，陳漢雄（筑波大学）

情報爆発の時代を迎え、情報統合の重要性は増加
しかし，一方で情報統合はますます困難に…

- 情報源の数と規模
- 情報源の異種性
- 情報源の動的変化，動的情報源

リソースマイニングに基づくアプローチ

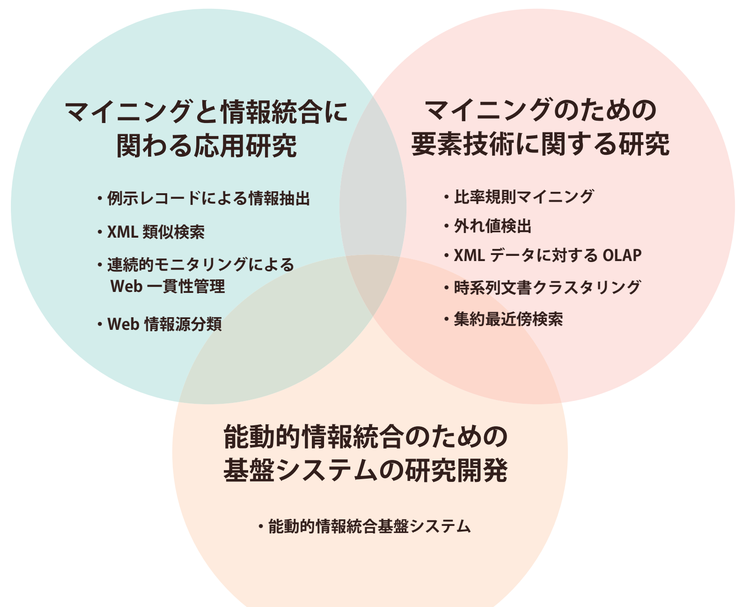
→ 知識発見と情報統合の融合



能動性
拡張性
分散環境への適応

- 統合対象の発見：情報源マイニング
- 動的変化：連続的マイニング
- 様々な情報源：異種データマイニング
- 複数情報源：クロスリソースマイニング
  → 情報源統合の高度化

主要研究テーマ



外れ値トランザクション検出

トランザクションデータの規則性から大きく逸脱したトランザクションを検出する枠組みの提案

- ▶ 高確信度の相関ルールに基づく外れ値度の導入
- ▶ 検出アルゴリズムの提案

外れ値度

トランザクションの希少性 = アイテム集合と共起すべきアイテムが存在しない

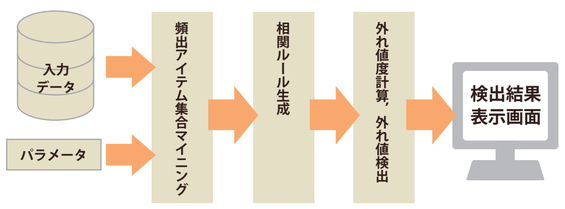
入力データ D 中のあるトランザクション t

D において高確率で成り立つ相関ルール

{ベーコン，ジャム，牛乳} → 比較
- {ベーコン，牛乳}→{卵}… {卵} が t には存在しない
- {ジャム}→{食パン}… {食パン} が t には存在しない

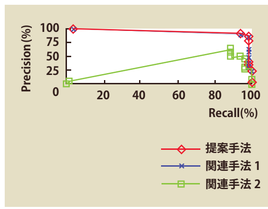
トランザクション t の外れ値度 $= \dfrac{|\{卵，食パン\}|}{|\{ベーコン，ジャム，牛乳\} \cup \{卵，食パン\}|} = 2/5$

検出アルゴリズム



検出精度

ネットワークアクセスログ [UCI MLR] から不正侵入の記録を外れ値として検出．その際の Recall, Precision を測定した．
提案手法は最大で 93.5% の F 値を記録．関連手法 1 と同等以上，関連手法 2 を上回る検出性能を確認した．



比率規則マイニングに関する研究

データ中の比率規則の抽出

- ▶ サポートと確信度の概念を導入
- ▶ 複数の線形関係が混在したり，一部分でのみ線形関係が成り立つ場合でも抽出可能

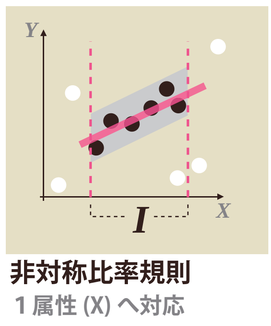
本研究における比率規則

- ▶ 線分とその近傍で比率規則を表現
- ▶ 比率規則に従うタプルの割合を用いてサポートと確信度を定義

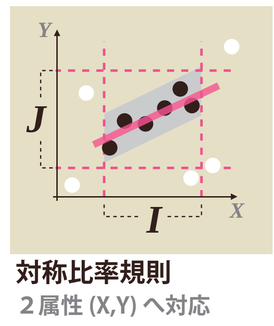
サポート … 分析領域のタプル中，比率規則に従うものの割合

確信度 … タプル全体に対する割合

サポート・確信度を最大とする最適比率規則を抽出

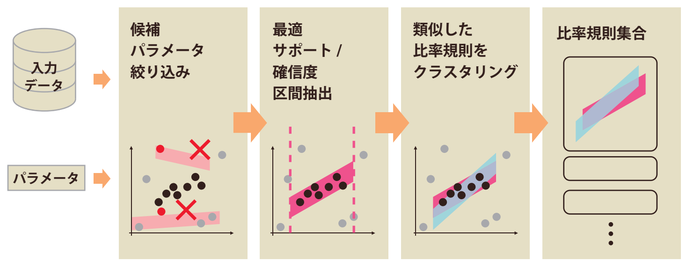


非対称比率規則
1 属性 (X) へ対応

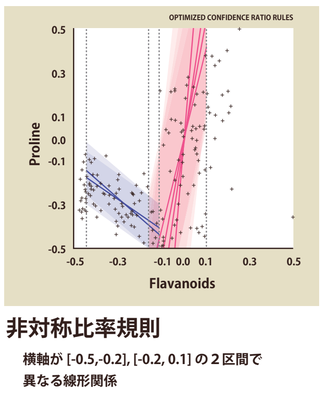


対称比率規則
2 属性 (X,Y) へ対応

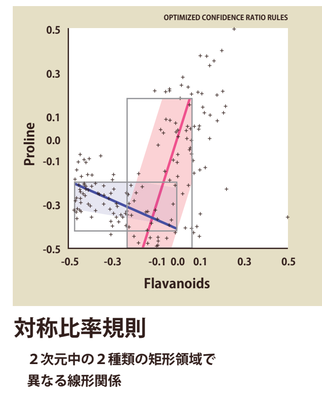
提案手法の概略



実験結果（ワインデータ）



非対称比率規則
横軸が [-0.5,-0.2], [-0.2, 0.1] の 2 区間で異なる線形関係

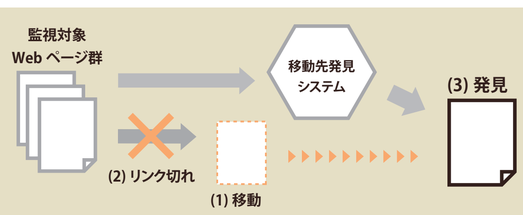


対称比率規則
2 次元中の 2 種類の矩形領域で異なる線形関係

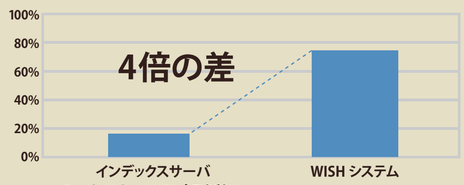
連続モニタリングによるWebコンテンツ一貫性管理

Web コンテンツ一貫性維持のための ページ移動先探索に関する研究

- ▶ Web ページの移動により生じるリンク切れの問題に着目
- ▶ ロボットにより Web ページ群を監視し，リンク切れを発見したときに Web ページの移動先を探索

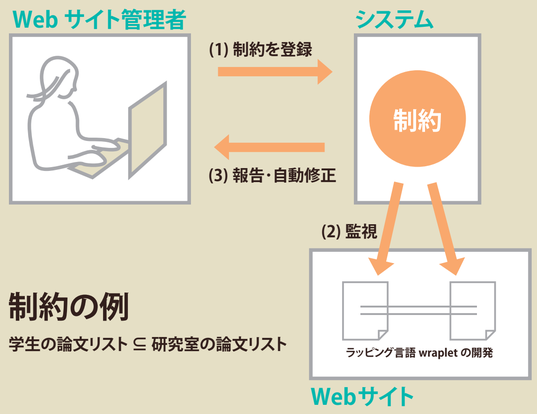


▶ 12 万リンクを対象とした実験結果



明示的な制約を利用した分散管理 Web コンテンツの一貫性維持

- ▶ 既存の Web コンテンツに対して後付けでコンテンツ一貫性管理を実現可能
- ▶ ラッピング・制約指定支援技術の開発



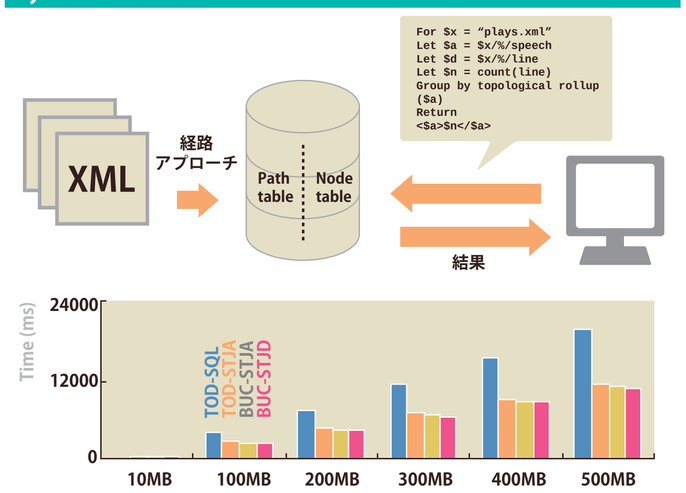
制約の例

学生の論文リスト ⊆ 研究室の論文リスト

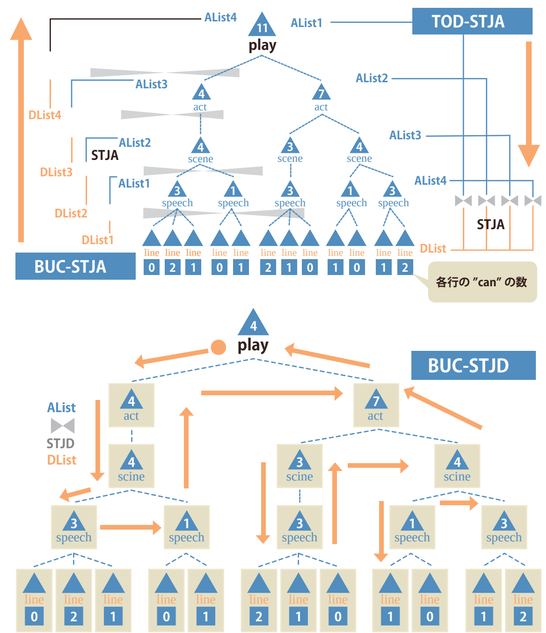
XML-OLAP XMLデータの多次元分析

- ▶ XML はデータ流通とストレージに広く使われているが，解析処理はまだ行われていない
- ▶ XML-OLAP 技術の構築に挑む
  → TOPOLOGICAL ROLLUP を Stack Tree Join を用いたアプローチで効率的に実現
  - ·Top Down (TOD-STJA)
  - ·Bottom-up (BUC-STJA)
  - ·Bottom-up (BUC-STJD)

System Overview



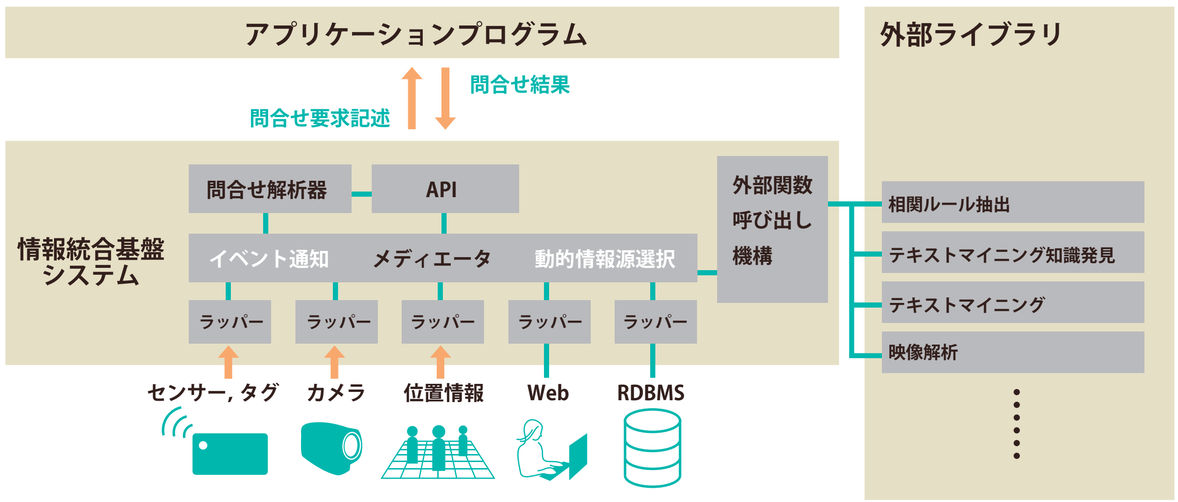
TOPOLOGICAL ROLLUP により "can" をカウント



能動的情報統合基盤システム

ストリーム等を含めた情報統合基盤：StreamSpinner

- ▶ データ到着やタイマーに連動し，イベント駆動で能動的に各種統合処理を実行
- ▶ 外部関数呼び出し機構やアプリケーション記述のための Java API による拡張性
- ▶ 膨大な情報源の中から，利用者の興味に応じて接続対象を動的に選択可能



広域分散データ処理技術

- ▶ 分散したストリームデータを効率的に収集・処理
- ▶ ネットワーク上の複数の StreamSpinner を連携させた分散処理の管理システムを開発

プログラム記述

分散問合せ

```
MASTER  chiba1
SELECT    *
  FROM  chiba [1s], hongo [1s], okubo [1s],
        kyoto [1s], imade [1s]
 WHERE  chiba.temp > 50  AND hongo.temp > 50
```

出力ストリーム定義

```
CREATE STREAM  Alert (ID long, Temp long)
```

アプリ固有の処理 (Java で記述)

温度データの受取り，火災検出処理，結果の配信

登録

管理画面

分散ストリーム処理環境
管理システム

管理・運用

分散ストリーム処理環境