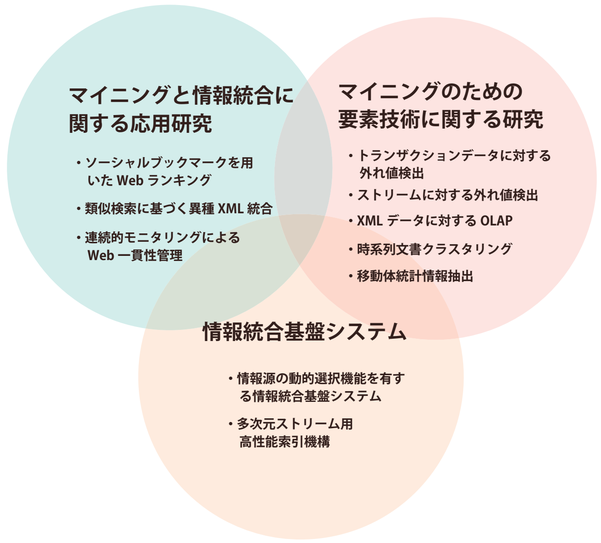
A01-03 能動的リソースマイニングに基づく異種情報統合基盤の研究

研究代表者：北川博之（筑波大学）　分担者：天笠俊之，森嶋厚行，川島英之（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

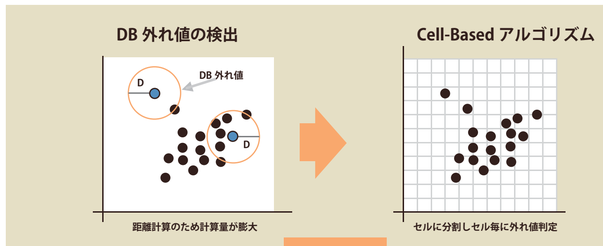
下記の３つの視点より研究を推進



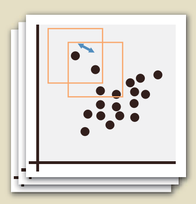
ストリームに対する外れ値検出

DEWS'08 優秀論文賞
DEXA'08 採録

▶ 目的：データストリームの特徴を活かして効率的に外れ値の検出を行う
▶ 手法：DB 外れ値を検出する Cell-Based アルゴリズム (E.M.Knorr) をもとにした差分処理を行う手法



提案手法：データストリームに対する外れ値検出アルゴリズム



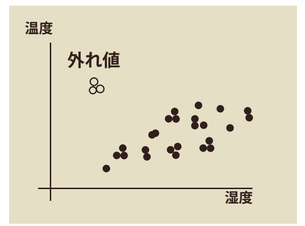
状態が変化したオブジェクトの影響及ぶ範囲に
対してのみ再外れ値判定処理

計算量の削減

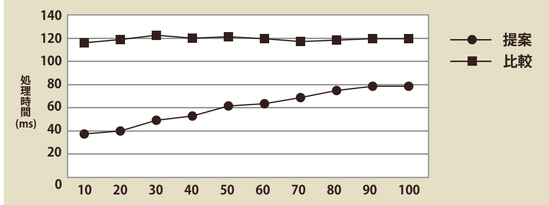
例：センサの観測値

▶ データストリーム特徴

データの分布が直前時刻のものと
類似する場合が多い



実験結果　移動体の割合の変化



外れ値トランザクション検出

WAIM'08 最優秀学生論文賞

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

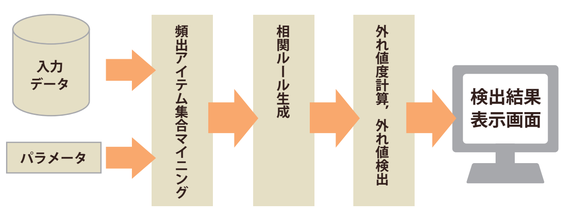
入力データ D 中のあるトランザクション t：{ベーコン，ジャム，牛乳}

比較

D において高確率で成り立つ相関ルール：
- {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

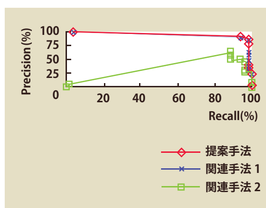
$$\text{トランザクション } t \text{ の外れ値度} = \frac{|\{\text{卵，食パン}\}|}{|\{\text{ベーコン，ジャム，牛乳}\} \cup \{\text{卵，食パン}\}|} = 2/5$$

検出アルゴリズム



検出精度

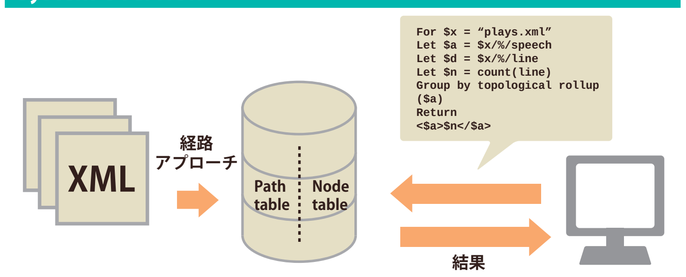
ネットワークアクセスログ [UCI MLR] から不正侵入の記録を外れ値として検出．その際の Recall, Precision を測定した．
提案手法は最大で 93.5% の F 値を記録．関連手法 1 と同等以上，関連手法 2 を上回る検出性能を確認した．

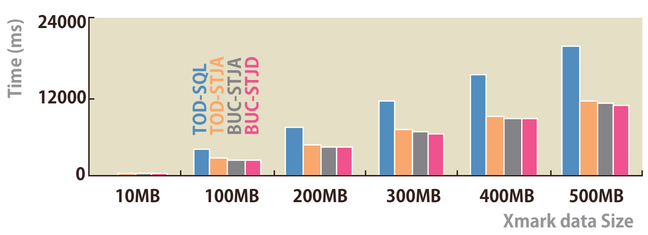


XMLデータの多次元分析

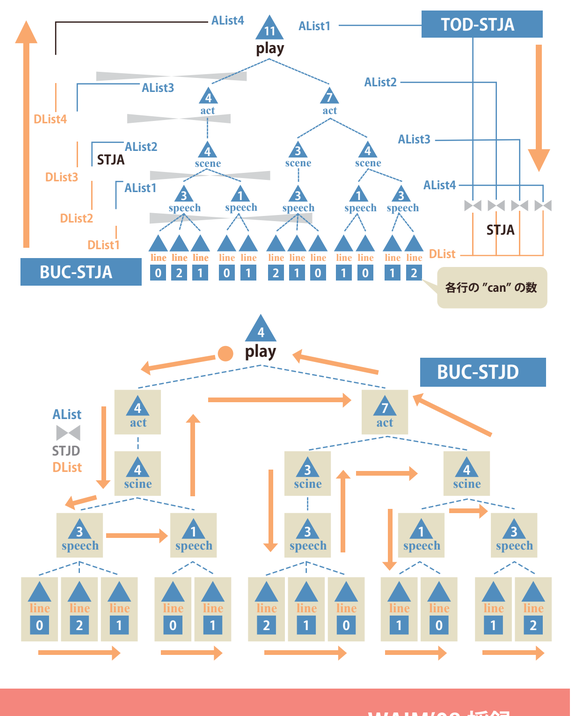
▶ XML はデータ流通とストレージに広く使われているが，解析処理はまだ行われていない
▶ XML-OLAP 技術の構築に挑む
→ TOPOLOGICAL ROLLUP を Stack Tree Join を用いたアプローチで効率的に実現
- Top Down (TOD-STJA)
- Bottom-up (BUC-STJA)
- Bottom-up (BUC-STJD)

System Overview





TOPOLOGICAL ROLLUP により "can" をカウント

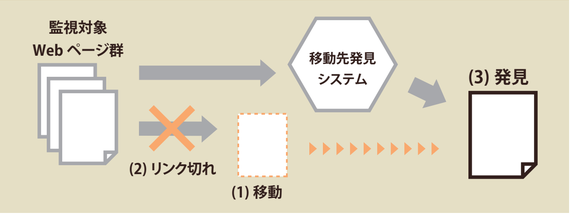


Webコンテンツ一貫性管理

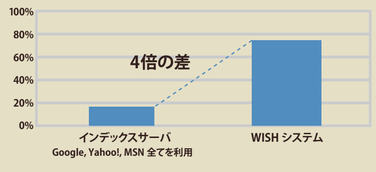
IEEE ICDE'08 採録

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

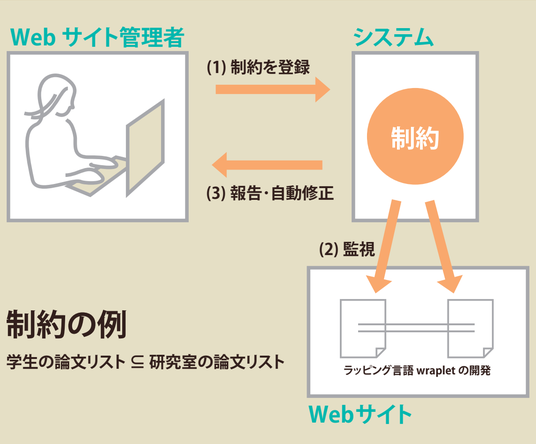


▶ 12 万リンクを対象とした実験結果



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

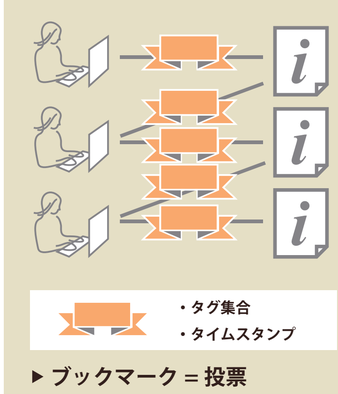
SBMを利用したランキング

WAIM'08 採録
DASFAA'09 採録

ソーシャルブックマーク (SBM) ユーザの興味・評価，専門度 (Hub 度) を反映させたランキング

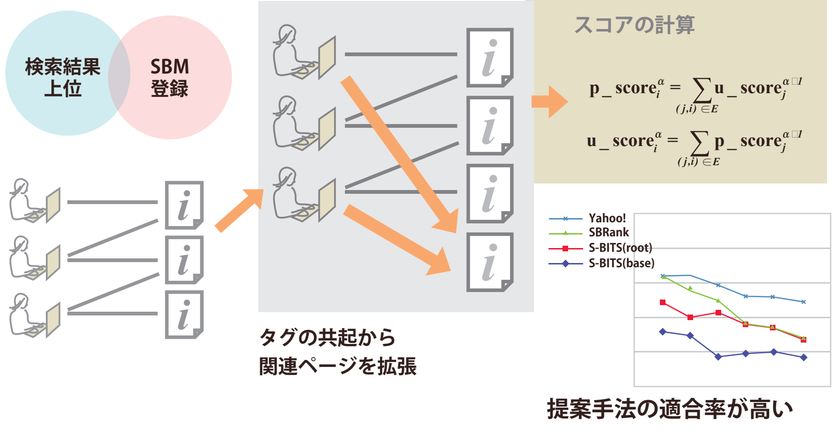
▶ 良いページは多くの良いユーザに参照される
▶ 良いユーザは多くの良いページを参照する

ソーシャルブックマーク



▶ ブックマーク = 投票
▶ ユーザ毎に異なる興・専門度

提案手法：S-BITS



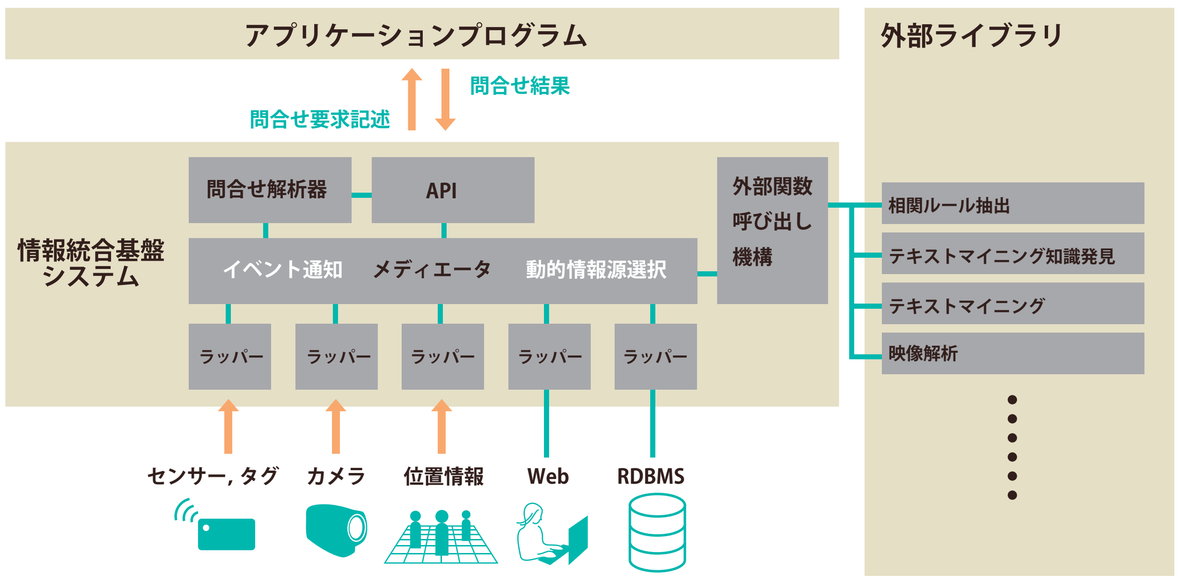
提案手法の適合率が高い

能動的情報統合基盤システム

Information Systems 条件付採録

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



更新に最適化した空間索引

IEEE Trans. on Knowledge & Data Eng. 採録

Rsb-tree: 準バルクロードを用いる R-tree

▶ 移動体のように頻繁に値が変わるデータへの索引機構の提案
▶ 準バルクロード：更新をメモリでバッファしてディスク転送コストを削減
▶ バッファサイズ，ページサイズ，更新 / 問合せ比率を変えて実験し，性能向上を確認
▶ I/O を要する演算を遅延 – split, remove, overflow treatment, and underflow treatment