【ICDE2014 勉強会】

A General Algorithm for Subtree Similarity-Search

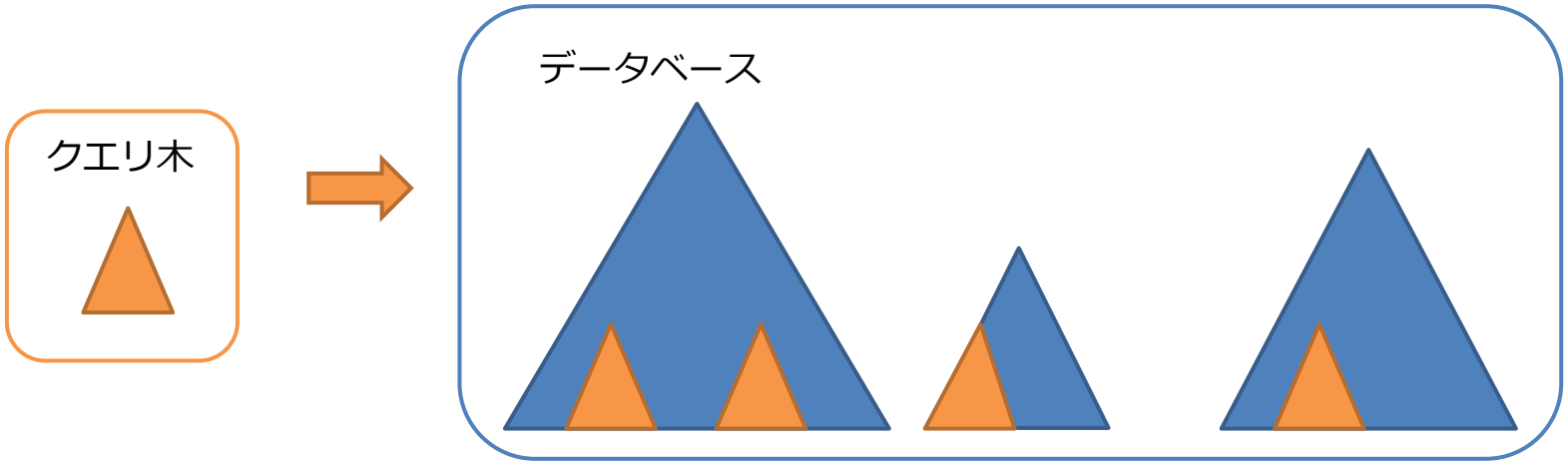
Sara Cohen, Nerya Or

Session 26: XML and Tree Data　担当 : 小柳 涼介（筑波大学）

A General Algorithm for Subtree Similarity-Search

Sara Cohen, Nerya Or

- 問題：類似部分木検索
  - 与えられたクエリ木に類似した部分木をデータベースから検索
- 応用
  - RNAの二次構造比較，自然言語処理，文字認識，半構造データの比較，等
- 本手法の特徴
  - 様々な木と木における類似度尺度が適用可能
  - 動的計画法によるほぼ線形時間で実行可能なアルゴリズム



A General Algorithm for Subtree Similarity-Search

Sara Cohen, Nerya Or

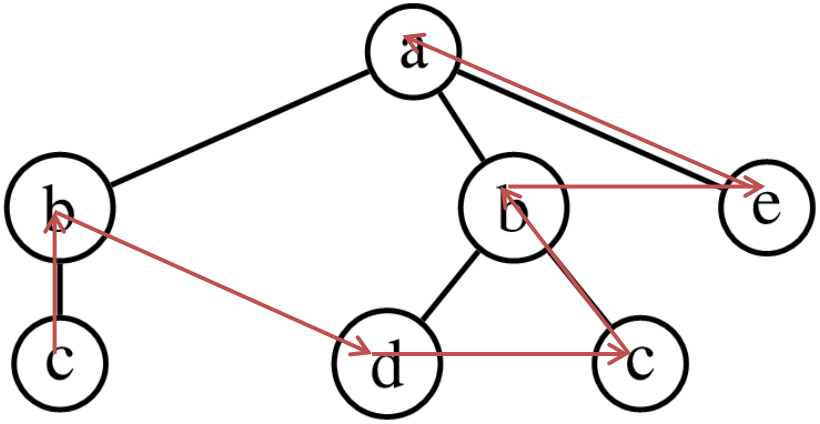
- フレームワーク
  - 木と木の距離（類似度）を二つの関数で表現
    - 木をprofile (multi-set) に変換する関数 → Tree Profile Functions
      - 例: 各ノードが持つラベルの集合
    - 二つの multi-set の距離を求める関数 → Multiset-distance function
      - 集合の要素数$|M_1|, |M_2|$と共通要素の数$|M_1 \cap M_2|$から計算
      - 例: Dice 関数，Jaccard 係数
  - 過去に考案された多くの木と木の距離指標がこのフレームワークに適用可能 → General !
    - pq-gram Distance
    - Windowed pq-gram Distance
    - Binary Branch Distance
      - 適用可能
    - Tree Edit Distance
      - 適用不可能

A General Algorithm for Subtree Similarity-Search

Sara Cohen, Nerya Or

- アルゴリズム
  - データベースの Profile は予め計算しておく
  - データベース内の木を後置順に走査し、
  - 子ノードのクエリとの共通要素 から
    親ノードのクエリとの共通要素を計算（動的計画法）
  - 類似度（距離の近さ）の上位 k 件のみを保持

- 計算量
  - おおよそ $O(|T|\log|k|)$
    - $|T|$はデータベースに含まれるノード数
    - ※$|T|$が巨大である場合
    - クエリサイズの影響が小さい

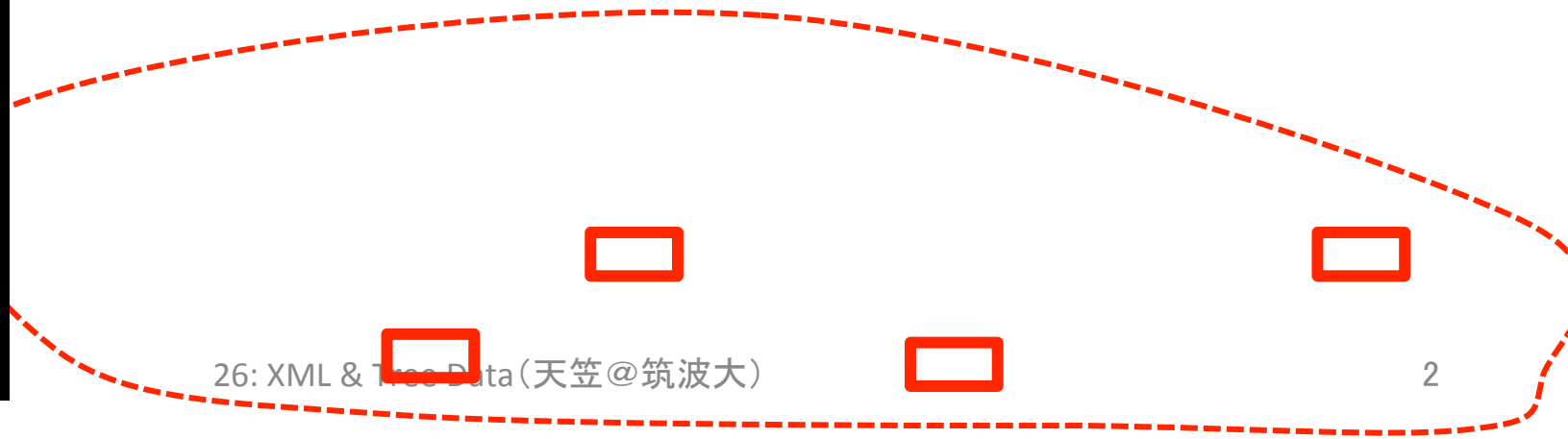


Breaking out of the MisMatch trap

Y. Zeng, Z. Bao, T. W. Ling, H. V. Jagadish, G. Li

- 問題
  - データベース中に期待する結果がないときでも，無関係な結果が（大量に）返却される.
    - MisMatch問題
  - XMLキーワード検索におけるMisMatch問題への対応.
- 技術的な貢献
  - XMLキーワード検索におけるMisMatchの検出.
  - データ駆動による，結果の説明と問合せ候補の生成.
  - ビットマップベースのノードラベルによる高速処理.

- 赤い（Red）Vaio Wの値段（price）を知りたい場合，該当する商品がないため，キーワード検索の結果がshop要素になってしまう.
- そうなった原因となるキーワードと代替候補を示したい.



Breaking out of the MisMatch trap

Y. Zeng, Z. Bao, T. W. Ling, H. V. Jagadish, G. Li

- MisMatch検出
  - 検索キーワードの型（type）から，結果の型（Target Node Type; TNT）を推定．検索結果の型とあわない場合にMisMatchと判定．
    - Vaio: online_mall/electronics/shop/laptop/model
    - W: online_mall/electronics/shop/laptop/model
    - red: online_mall/electronics/shop/laptop/color
    - price: online_mall/electronics/shop/laptop/price

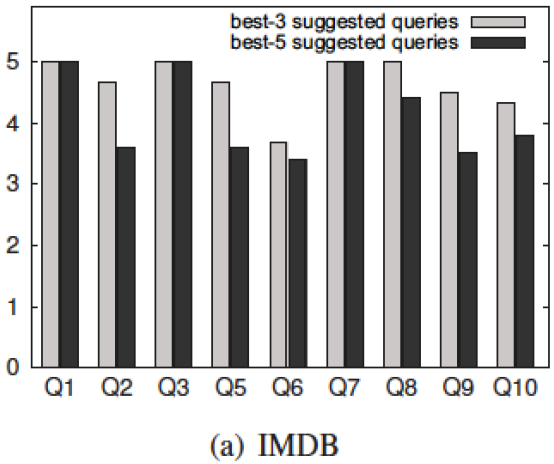
    ➔ “shop” と整合しないためMisMatchと判定．
- 原因説明と問合せの推薦
  - 問合せキーワードの中から，MisMatchの原因となったキーワードを特定．
  - 識別性（distinguishability）を提案．tf-idfの考え方に基づき，あるタイプのノードに少数しか含まれないキーワードに対して，値が大きくなるよう設計．
  - 各キーワードに対する型tにおける識別性が閾値以上の場合，TNTを計算．それに基づき近似解を計算．
- 上記の処理を効率的に行うためのノードラベルについても議論．

Breaking out of the MisMatch trap

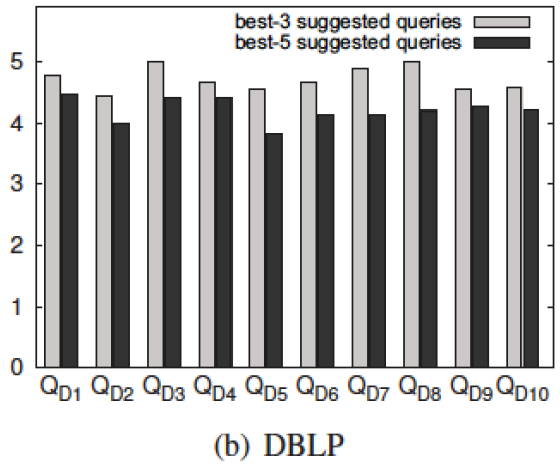
Y. Zeng, Z. Bao, T. W. Ling, H. V. Jagadish, G. Li

- 評価実験
  - データセット：IMDB, DBLP, IEEE Pub.（INEX）
  - サンプル問合せ（→）
- Ground truth
  - 15人の被験者により，過半数がMisMatchありと判定した問合せをMisMatchありの問合せと判定.
- 推薦問合せの平均品質（→）

IMDB:90MB			
#	Query	suggested queries	best-3 suggested queries (Format: explanation → suggested options)
Q1	Gladiator Spanish	5	(language): Spanish → English / Japanese / French
Q2	Spielberg DiCaprio Action movie	6	(genres): Action → Biography / Crime / Drama
Q3	Neo hacker phonebooth	3061	(keyword): phonebooth → computer / software / programmer
Q4	Warner Bros. movie	0	None
Q5	Italy Betty Fisher	12	(country): Italy → France / Canada / USA
Q6	Spielberg Schwarzenegger	58	(name): Schwarzenegger→ Meredith Brooks / Jim Conroy / Dean Spunt
Q7	Terminator 3 cast Sarah	19	(name): Sarah → Nick Stahl / Claire Danes / Kristanna Loken
Q8	Panic Room 2001	11	(year): 2001 → 2002 (title): Panic Room → Promised Land / Nowhere Road
Q9	Ettore The Man movie	1189	(director): Ettore → Ethan Coen / Salvatore Maira / Massimo Sani
Q10	boy death ghost love	992	(keyword): love → orphanage / bully / bomb



(a) IMDB



(b) DBLP

XQuery streaming by Forest Transducers

S. Hakuta, S. Maneth, K. Nakano, H. Iwasaki

- 研究の目的
  - Macro Forest Transducer（MFT）を利用した，XMLストリームに対するXQuery問合せ処理方式の提案.
- 提案手法のポイント
  - Macro Forest Transducer: 森から森への変換を，森の上の相互再帰関数として表現.
  - XQuery問合せから，ストリーム処理可能なMFTを生成.
    - GCXでサポートされているXQueryのクラスより大きいクラスの問合せをサポート.
  - MFTの合成について議論.
    - あるクラスのMFTであれば，合成可能であることを示した．MFTを合成することにより，中間結果を削除.
  - 最適化手法を提案.
    - パラメータ削除，stay moves（ルール右辺での関数呼び出し）の削除.
  - 実験による性能評価.
    - 最適化により，10倍程度の高速化を達成.