【VLDB2011勉強会】

Session 22: Data Integration

担当：田島 敬史（京都大学）

Online Data Fusion

by Xuan Liu (NUS), Xin Luna Dong (AT&T Labs) et al.

- Web上には多くの相反する情報や誤った情報
- しかも，それらはコピーされ広がる

  複数の情報源からの情報を集約することで正しい情報を提供

- しかも，それをオンラインに
  - 最初の情報源からの情報を得たらすぐに表示開始
  - 新しい情報源を調べる度に，予想される正解と確信度を更新
  - これ以上，他を見ても結果が変わらないという時点で終了

課題

- 未取得の情報源やコピー関係も考慮した確信度の計算
- 情報源を調べる順番：
  - 早く正解に達するように
  - 早く打ち切れるように

コピー関係がなければ単に信頼度順

前提 以下が既知とする

- $\alpha(S)$ — 情報源 S の情報が正しい確率
- $\rho(S \to S')$ — S の情報が S' と同じ時にコピーである確率

提案手法の概要

1. コピーである確率，残りの情報源がどんな答かを考慮して，全ての可能世界とその確率を考える
2. 各時点で，各候補値 v について以下を求める
   - 最大確率
   - 最小確率
   - 期待確率
3. 「vの最小確率 $>$ 他の値の最大確率なら打ち切り」は厳しすぎ実験によると以下の打ち切り条件でも十分
   - vの期待確率 $>$ 他の値の最大確率なら打ち切り
   - vの最小確率 $>$ 他の値の期待確率なら打ち切り

問題点と本論文での解決策

- コピー元が未調査でコピー先と同じ答か不明の場合は？
  - **conservative** 手法：コピーであると仮定しておく
  - **pragmatic** 手法：コピーでないと仮定しておく
- 最大・最小確率の正確な計算は **PTIME** なので近似

  v の最小確率の近似 (最大確率も同様)
  - v と言っている調査済情報源は,「実は未調査の情報源のコピー」と思った方が点が下がる場合はそう仮定
  - v 以外を答えている調査済情報源は，全てコピーでないと仮定
  - 未調査情報源は，現時点で v 以外で最上位の値を答えると仮定
- 情報源を調べる順番は？
  - **conservative** 手法：情報源が与えうる最小の貢献 (コピーだと仮定) の降順
  - **pragmatic** 手法：情報源がコピーでない場合の貢献の降順. コピーが判明する度に貢献を計算し直すので複雑だが高性能

Synthesizing Products for Online Catalogs

by Hoa Nguyen (U. Utah), Ariel Fuxman (Miscrosoft) et al.

目的 価格.com のような統合カタログの自動生成

- 統合カタログは各製品カテゴリ (例：HDD) 毎にスキーマがある
- 各販売店も各製品カテゴリ毎に商品情報のスキーマを持つ

各販売店サイトの商品情報を統合し統合カタログを生成

1. 実体間の対応付け：各販売店の商品情報を「製品」に対応付け
2. 属性間の対応付け：販売店毎に情報中の属性数，属性名が違う

属性間の対応付け

- 従来のスキーママッピング：
  現れる属性値の分布の類似度で属性を対応付け
- 同じ属性でも商品情報での分布と製品情報での分布は異なる
  → 製品のうち過去に商品情報とマッチしたものの分布を使う
- 分布を見る際，ある販売店のあるカテゴリのデータが少数かも
  → 分布を作るためのグルーピングに以下の三つを併用
  - 各販売店，各カテゴリ毎に属性値の分布を生成
  - 全販売店のそのカテゴリのデータ中の同じ名の属性を使う
  - その販売店の全カテゴリのデータ中の同じ名の属性を使う
- 分布類似度は **Jensen-Shannon** 情報量と **Jaccard** 係数を併用
- 「**3**種類のグルーピング」*「**JS**と**Jaccard**」の**6**つの属性に基づく「同じ意味の属性か？」の分類器を生成．個別の属性を用いた実験より性能改善.

- 学習用データは？ → 自動生成
  - ある販売店・カテゴリの商品情報と同じカテゴリの製品情報に同じ属性名 → 正例
  - 正例となる属性と同じ情報内にある他の属性 → 負例

実体間の対応付け

- 属性間の対応付けで「**Model Part Number**」か「**UPC**」に対応づけられた属性の属性値に基づいて対応付け