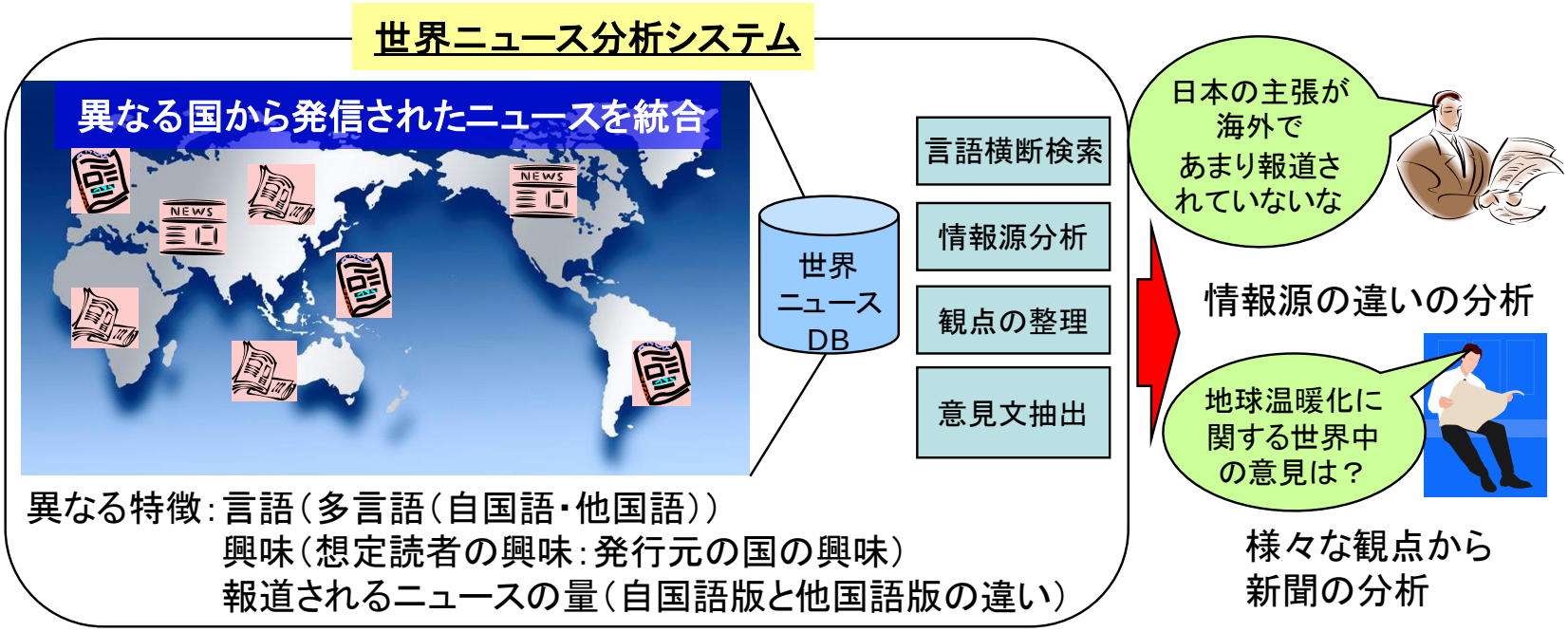
VSOPによるデータベースのマルチファセット分析に関する検討

吉岡真治（北海道大学）

yoshioka@ist.hokudai.ac.jp

背景：複数国のニュースサイトの分析

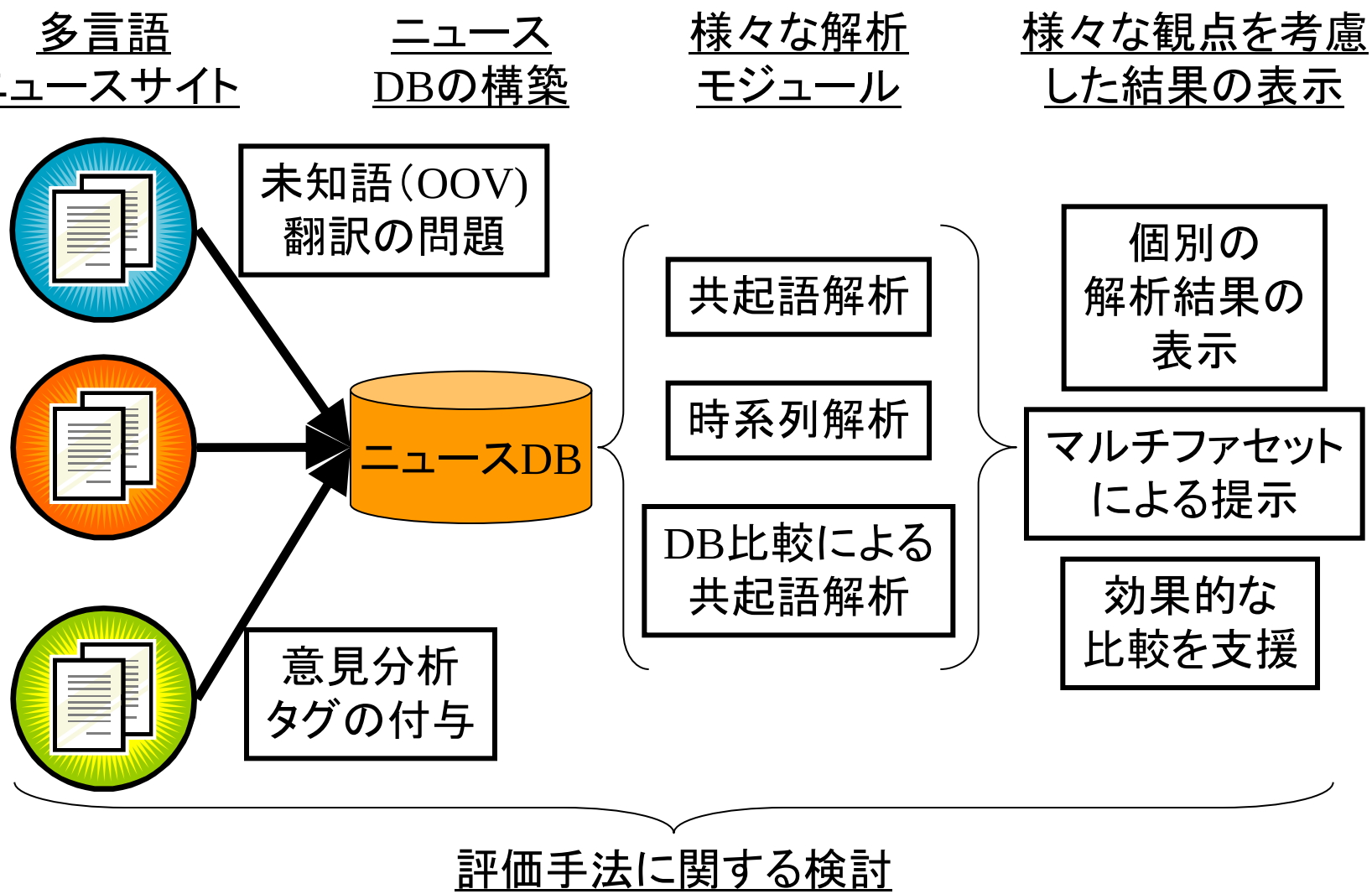
- 地域ごとに取り上げるニュースが違う
  - 国連の会議のような全世界的に興味をもたれる事象の報道内容は、地域（読者）の興味の影響を受ける
- 複数ニュースサイトの比較による世界ニュース分析システムの提案



本研究の目的

- 複数国のニュースサイトの比較手法の提案
  - 各国が関心を持つトピックの比較分析
    - 各国におけるメジャーなトピックの違い
      - バースト解析の比較
      - 相関性の高い共起語による関連語の抽出
    - 各国の比較によるトピックの分析
      - 相関性の比に注目した関連語の抽出
        マイナーメジャーな関連語や無視されている関連語を抽出
  - 様々な観点による分析
    - 意見分析の結果を考慮した比較
      - 意見の賛否を考慮した比較
    - マルチファセットによる結果の提示

研究トピックとその関係



DB比較による共起語解析

- 各々のニュースサイトが興味を持っている項目
  - 世界的に注目を浴びる事象がおこった場合には、ニュースサイトごとの違いはあまり見出せない。
- 異なるテキスト群の比較による特徴語発見
  - 従来型の相関性の高い共起語ではない、特徴語を発見する手法の開発
  - 他の国のニュースサイトよりは興味を持たれている項目や、他の国のニュースサイトに比べて無視されている項目などに注目

ニュースサイトの比較解析の効用

- 他国の動向の分析
  - 政策立案などへの応用（例：EMM NewsExplorer)
- 新たな視点の発見
  - 他国の興味からニュースの新たな切り口に気づく。
- 各々の読者が事象を理解するために用いている背景情報の違い
  - 各々の国のマスメディアの報道の違いにより、読者が持っている情報が違う。
  - 基本となる背景情報が違うと、コミュニケーションがうまくいかない。

マルチファセットによる文書群の分析

- 文書を特徴づける様々なファセット
  - 情報源、単語、固有名詞、日時、場所、意見、…
- 文書群に対して、これらのファセットを利用することにより、以下の機能の実現を目指す。
  - ファセットごとでの絞り込み検索
    - ファセットごとに条件を与えて、複合的な検索条件により絞り込み検索
  - 対比
    - 対比させたい制約条件に対する結果を複数表示させることによって、その内容を比較
  - 可視化
    - ファセットごとの情報集約を行い可視化

マルチファセット検索の高速化

- 複雑なブーリアン検索と共起頻度情報の獲得
  - データベースをオンメモリ上に展開することによって高速化を目指す
    - ユーザが大まかに初期検索式を与えることによって、検索対象とする記事の絞り込み
      - 数十万記事のオーダーから数千記事のオーダーに
    - ZDD（Zeroサプレス型BDD)による記事情報のメモリ空間への展開と分析
      - トランザクションデータに対する高速な論理演算が可能

システムへの入力

- 入力
  - ファセット情報＋文書ベクトル
    - ファセット情報
      - 文書ベクトルの特定次元とファセットを対応させる
      - 例:
        - » 情報源:1～10
        - » 固有名詞:11～50
        - » 単語:51～800
        - » …
      - 文書IDと対応付けることにより、一つだけ連続量の情報を扱う
        - » 例:期間　3/20　1～20、3/21　21～40、…
    - 文書ベクトル
      - 各文書を上記に対応する高次元ベクトル(1,0)で表現

システムへの入力(具体化)

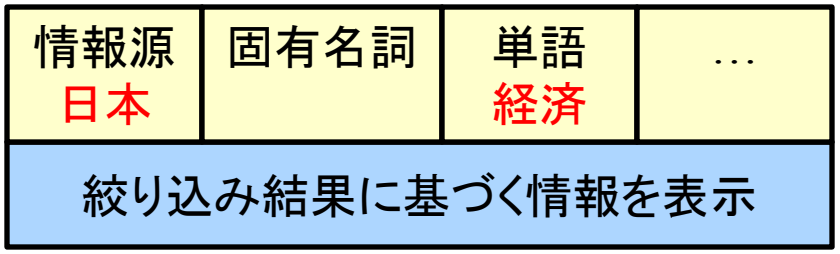
- ファセット情報(XML形式などで表現)
  - ファセットのタイプ名
    - 次元のメタ情報
      - 内容
      - 付加情報(DFなどのDBグローバルから得られる情報)
    - ファセット内の項目を選択するための手法
      - プルダウンリスト、パターンマッチ、…
- 文書情報
  - 文書ベクトル
  - メタ情報
    - 検索し、表示するための情報

入力データの作成

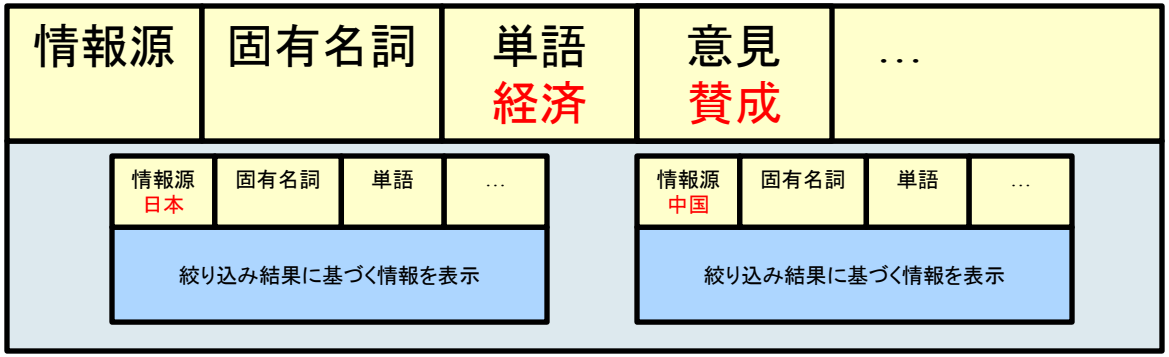
- 文書データの獲得
  - 現在構築中の新聞記事データベースに対し、一定の条件を与えることにより、1000件程度の記事を取得
    - どれくらいの件数までが良いかは、システムの応答性能次第
    - 意見分析などの場合は、記事単位ではなく、分単位というのも考えられる。
- ファセット情報と文書ベクトルの生成
  - 対象とするデータに関する情報から、スキーマ付きのデータをファセット情報＋文書ベクトルとして変換

出力のイメージ

- ファセット情報を用いたグローバルとローカルの絞り込み
  - ファセット情報を用いて、ファセットごとの検索条件を設定するインターフェースを構築



  - グローバルな制約とローカルな制約
    - 全体に関わる制約条件とローカルな制約を利用

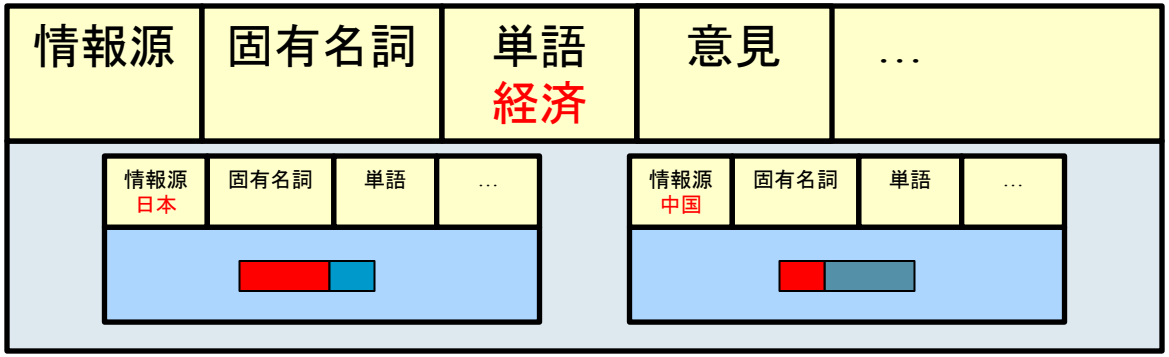


ファセットを考慮した情報の集約表示

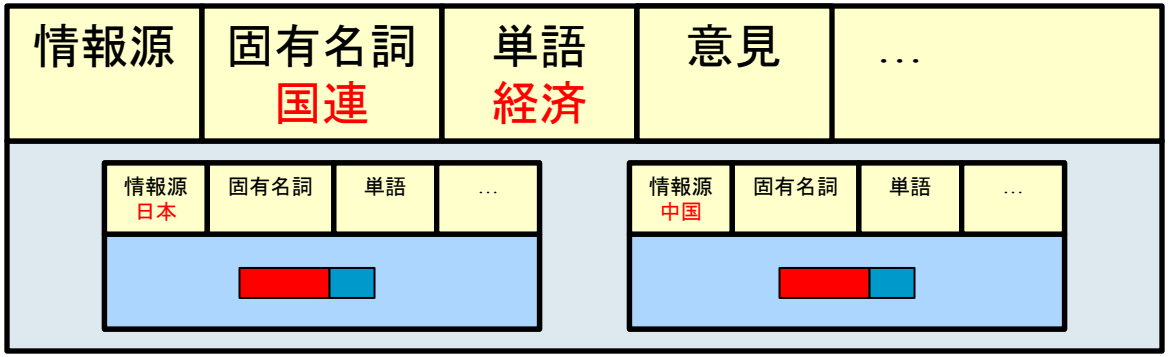
- 情報の表示方法の定義とファセットとの関連付け
  - 情報の表示方法（頻度情報を主に利用）
    - 百分率のグラフ、単語リスト、頻度順の棒グラフ、…
  - ユーザは、表示する内容と情報の集約方法を選択
    - ファセット＋表示方法の選択
      - 絞り込みのインターフェースと同等のものを利用した対応付け
    - 例：
      - 賛成・反対：百分率のグラフ
        - » 赤を賛成、青を反対として、グラフで表示
- ファセット情報の表示テンプレート
  - よく使う分析シナリオに応じた表示法を定義し、基本的には、その中から選択

インタラクティブなシステムの構築

- 利用イメージ
  - 経済という語を含む記事に関する日中比較



  - 制約を与えるとインタラクティブに内容をアップデート



簡単な例

- VSOPコマンドによるファセット情報の比較

vsop> source henoko2.vsop 「辺野古」を含む記事のVSOP表現
vsop> print (F/s0).TotalVal 情報源s0（読売）からの記事数
173
vsop> print (F/s1).TotalVal 情報源s1（朝日）からの記事数
172
vsop> G = (F/s0).FreqPatC(50) 情報源s0の頻出パターン
vsop> H = (F/s1).FreqPatC(50) 情報源s1の頻出パターン
vsop> print G.Permit(H) どちらの情報源でも共通の頻出パターン
vsop> print G - G.Permit(H) s0でのみの頻出パターン
vsop> print H - H.Permit(G) s1でのみの頻出パターン

結果

- 現状では、複雑すぎて、良く分からない。
- アイテム数2とか3に限定する代わりに、中頻度や低頻度のものを取り出せると、もう少し、理解しやすいものができるかもしれない。