【ICDE2014勉強会】

Session 10:Strings and Texts

担当：嫁兼弘修（同志社大学）

Session 10: Strings and Texts 論文リスト

- MassJoin: A MapReduce-based Method for Scarable String Similarity Joins
  - Dong Deng, Guoliang Li, Shuang Hao, Jiannan Wang, Jianhua Feng
  - MapReduceとlight-weightフィルタを利用し String Similarity Join の際の伝送コストを大幅に削減
- Efficient Instant-Fuzzy Search with Proximity Ranking
  - Inci Cetindil, Jamshid Esmaelnezhad, Taewoo Kim, Chien Li
  - クエリとして入力されたキーワードの分割方法とインデックスの作成方法を変更していい感じになりました（？）
  - 理解が困難なため断念

MassJoin: A MapReduce-based Method for Scarable String Similarity Joins

- 概要
  - String Similarity Joins
    - データ統合の際に2つの文字列コレクション内に存在する同じような文字列のペアを発見
    - 類似度の定量化には set-based(e.g. Jaccard) と character-based(ex. Edit distance)が主な手法として存在
  - MapReduceを基に拡張性のある String Similarity Joins を行おう！
    - 文字列ペアを全て処理してゆく従来の方法では効率が悪いので フィルタをかけて最終的に処理する数を減らす
    - フィルタをスルーしやすいシングルトークンも提案するlight-weightフィルタで削減
    - 類似度計算の際Jaccard係数などを使うと計算量が$O(l^3)$になるので$O(l)$まで縮小

実行時間を著しく減少

MassJoinフレームワーク

- MassJoinフレームワーク
  - 統合する2つの文字列コレクション R, S の文中の文字列 r , s に対し set-based と character-based の両方の相似関数を用い閾値を設定
  - r, s に対し類似度を計算，設定した範囲内に収まれば類似する文字列に対しシグネクチャを作成，文字列を値と設定するkey-value ペアを設定
  - key-value ペアのリストとして統合比較を行ってゆく

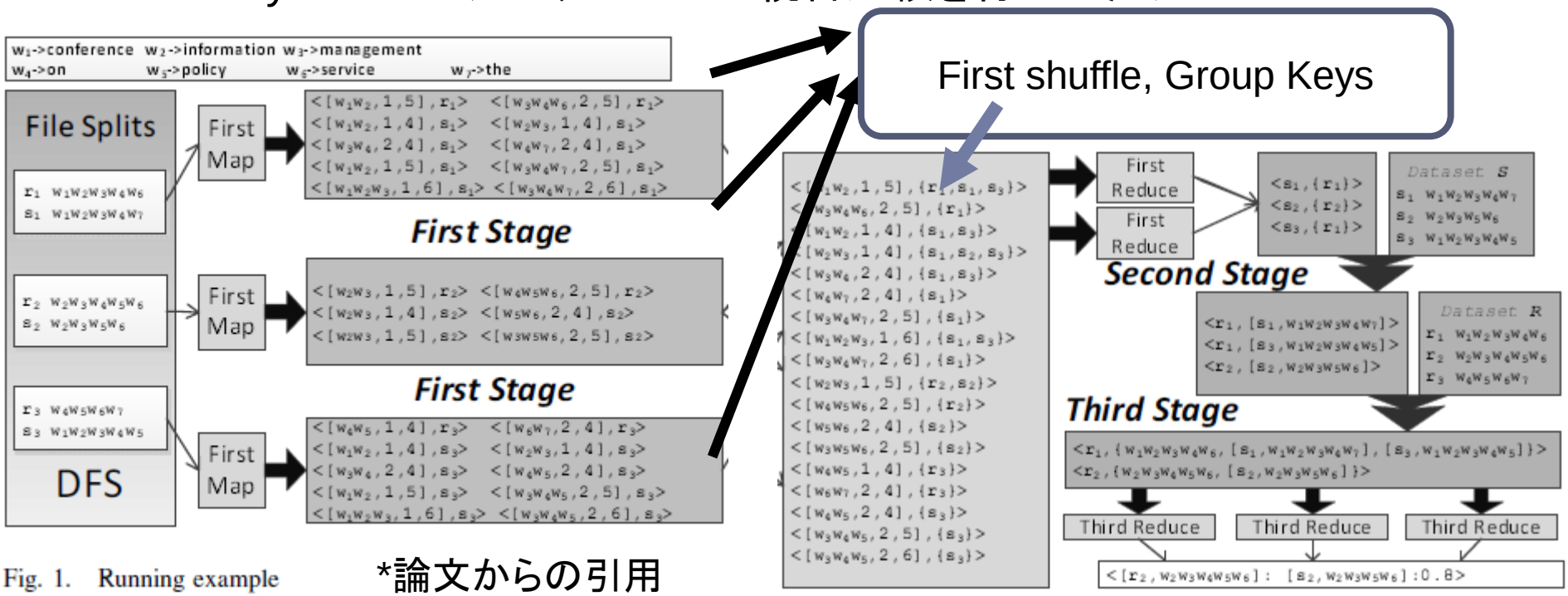

Fig. 1. Running example

*論文からの引用

light – weight フィルタ

- 概要
  - key-value ペアとして設定されたものの実際には類似度が低くなるような多くの文字列を削減
  - 文字列中のトークンを整数値に置換し，その整数値の集合を文字列に対するフィルタとして利用
    - 文字列中にトークンが多ければ類似度は高くなっても伝送コストが大きいままになる
  - character-based で類似度を計算する際に用い，設定した閾値を下回る key-value ペアを検出し除外

実験結果

- PrefixFilter との比較
  - 生成するkey-value pairが少なくてよい

実行時間を短縮

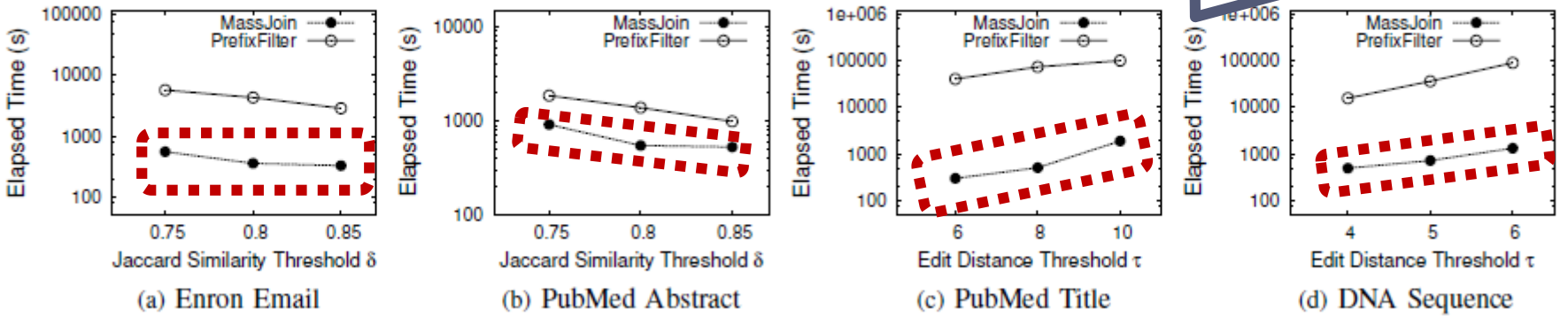

Fig. 4. Comparison with state-of-the-art methods(VSMARTJoin and FuzzyJoin are out of memory).

- ノード数を変化させた場合の処理速度

増加に伴い上昇

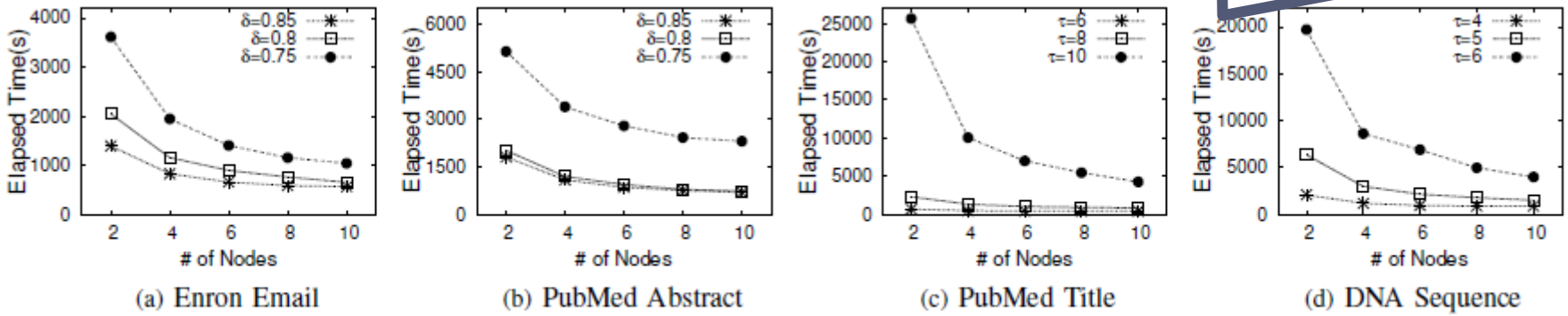

Fig. 5. Speedup by varying number of nodes.

*論文からの引用