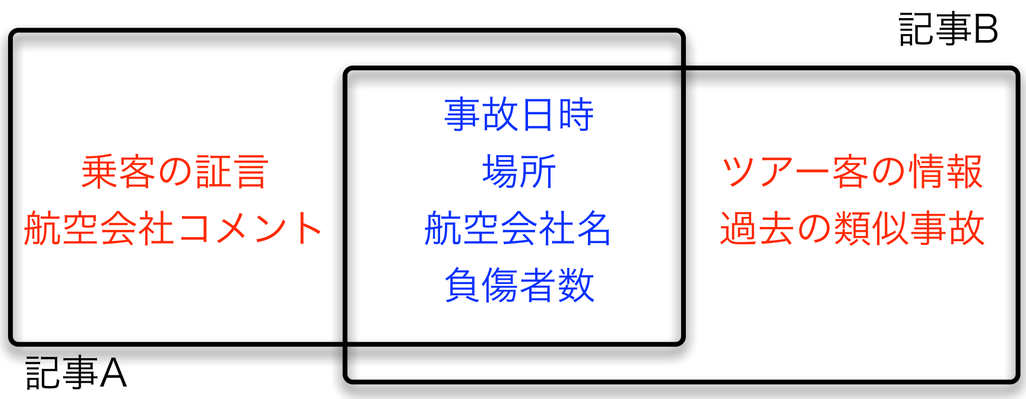
記事間の差異に着目した　ニュース閲覧システム

野呂智哉　　東京工業大学　大学院情報理工学研究科

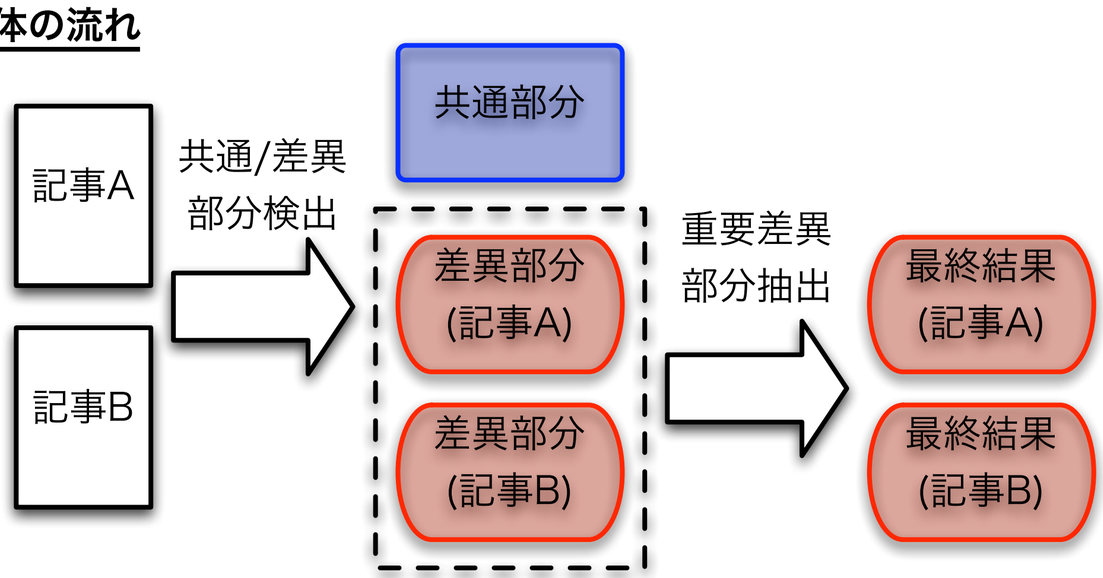
様々な機関がWeb上で大量のニュース記事を配信
　　効率よく閲覧できるようにしたい

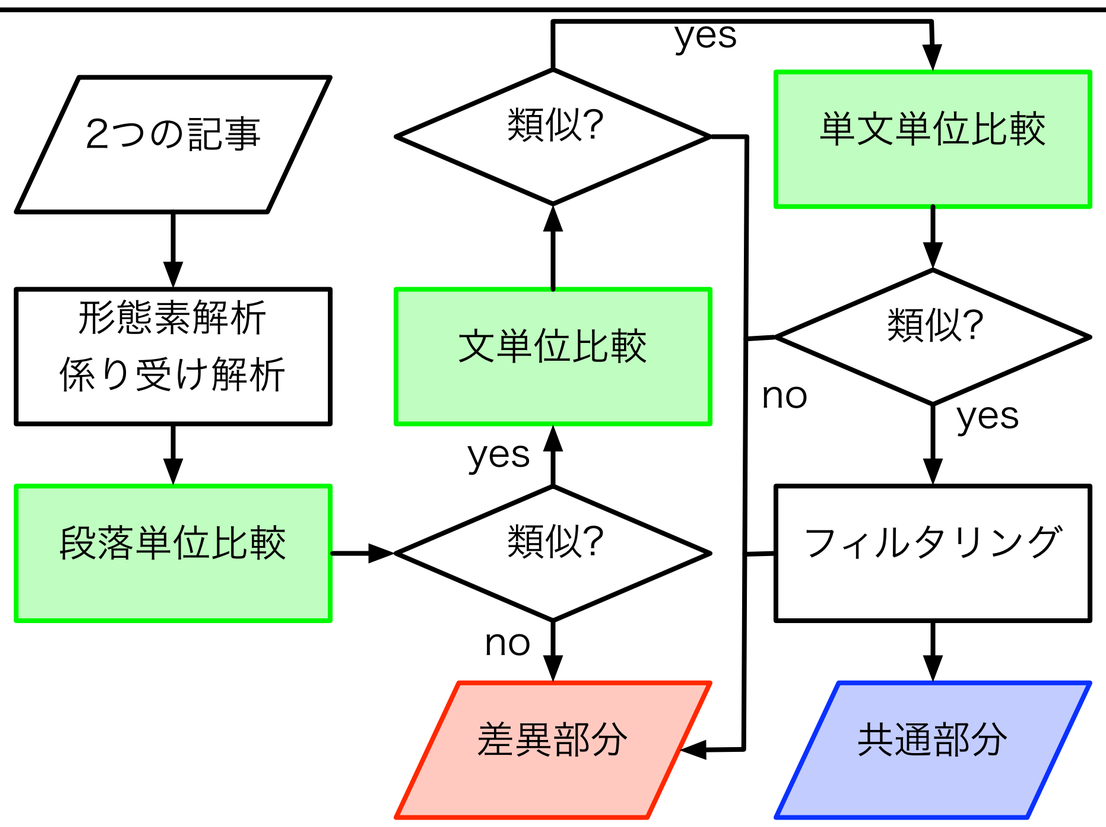
1つの方法: 類似記事ごとにクラスタリング，カテゴリ分類
　　(例) Google News，Yahoo! News
　　それだけで十分か?
　　　　Google News USは4500のサイトから記事を収集，
　　　　1トピックに数百から数千の記事

同一内容を扱う記事でも，完全に同じであることはない
　　(例) 航空機が乱気流に巻き込まれる事故



全体の流れ



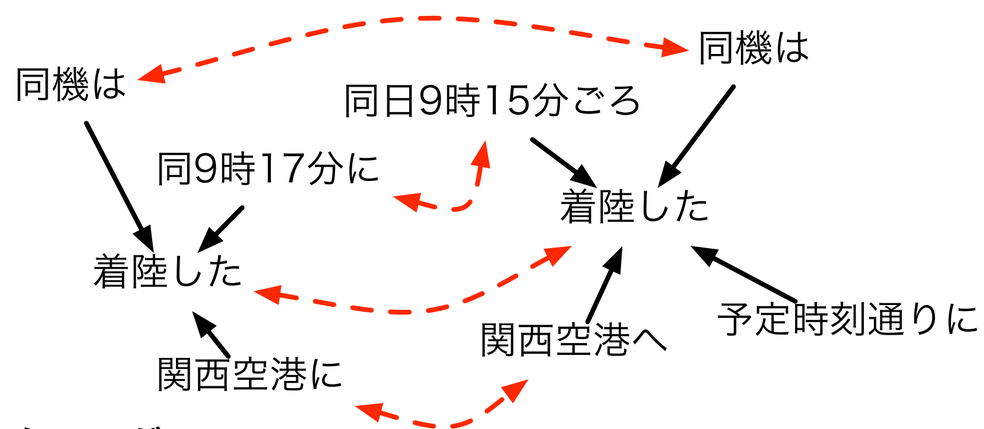


3つの比較単位
1. 段落単位
2. 文単位
3. 用言を中心とした係り受け構造単位 (単文単位)

形態素解析 / 係り受け解析
　　Juman + KNP，Mecab + CaboCha併用
　　KNPの結果のうち，並列・同格関係を採用
　　CaboChaの結果の中の固有表現の情報も利用

段落単位比較 / 文単位比較
　　内容語抽出，cosine類似度
　　固有表現は1語
　　数字は後続の名詞/接尾辞と結合
　　数字の違いを吸収 (DATE，TIMEを含む)
　　　　(例) 9人 ⇔ 10人，10時31分 ⇔ 10時30分
　　日本語WordNetを利用

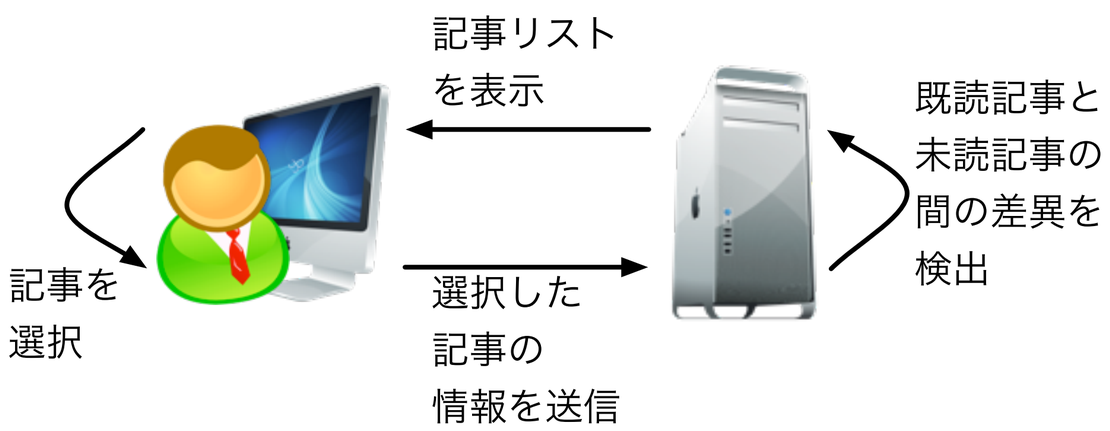
単文単位比較
　　係り受け構造も比較
　　助詞の情報も利用



フィルタリング
　　1つの単文が複数の単文と類似すると判定される可能性
　　スコア順にソート，上から順に採用，重複するものは除外

重要差異部分抽出
　　1つの差異部分(単文)Sに着目
　　　　1. Sに出現し，共通部分に出現しない語を取得
　　　　2. IDF値の上位N語の総和を計算
　　　　3. Sを含む段落ともう一方の記事の各段落との類似度がすべて閾値以下ならば，総和に一定数をかける
　　　　4. スコアが閾値以上ならば，重要差異部分と判定
　　IDF値はYahoo!ニュースの検索結果(記事数)をもとに計算

ニュース閲覧システム概要



　　複数の既読記事を連結，1つの記事として扱うことで，
　　2記事間の差異検出手法を適用
　　未読記事は，共通部分と差異部分の割合の積でソート

参考文献 (to be published)

Noro, T., Tokuda, T.: Detection of Difference between News Articles on the Same Topic Based on Sequential Comparison. Information Modelling and Knowledge Bases XXI, Frontiers in Artificial Intelligence and Applications, IOS Press. (2010)