ブログページ集合からのポスト及びコメントの自動抽出

NLP2010　PA1-39
吉田光男　乾孝司　山本幹雄
（筑波大学）

ブログページからポストとコメントを自動的に分離抽出する手法を提案します

□ 研究の背景・目的

■ ブログのコンテンツを有効活用したい
　◇ ポストとコメントを自動抽出する必要
■ ポストとコメントを分離する事で
　◇ コメントのみの研究利用を可能に
　◇ より正確なWebコンテンツ解析を可能に

□ 提案手法のアイデア・抽出例

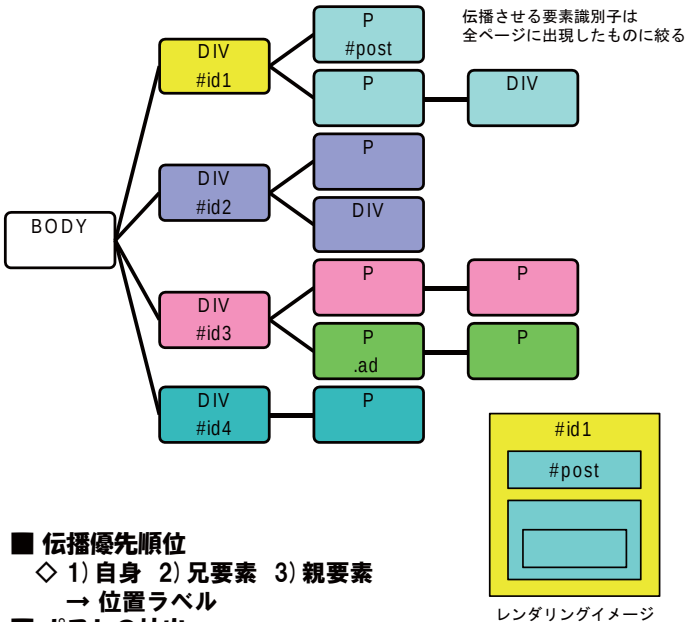
■ 部分の名前付け(グループ化)が必要
　◇ 要素識別子を伝播してグループ化
■ コンテンツ抽出
　◇ Webページ集合を用いて自動抽出(*)
■ コンテンツの再抽出
　◇ 部分の名前と要素名をもとに再抽出

ポスト/コメント自動分離抽出例

(*) 吉田光男, 山本幹雄. 教師情報を必要としないニュースページ群からのコンテンツ自動抽出. 日本データベース学会論文誌, Vol.8, No.1, pp.29-34, 2009.

□ アルゴリズム・実験結果

要素識別子の伝播イメージ



■ 伝播優先順位
　◇ 1) 自身　2) 兄要素　3) 親要素
　　→ 位置ラベル
■ ポストの抽出
　◇ 全てのページで
　　コンテンツとなった位置ラベルの付与されたコンテンツ
■ コメントの抽出
　◇ ポストにならなかったコンテンツ

コンテンツの再抽出

■ 従来手法の欠点
　◇ 連続する
　　コンテンツを取りこぼす
■ 再抽出条件
　◇ 位置ラベルと要素名が
　　抽出成功コンテンツと同じ
■ 位置ラベルの特徴
　◇ 内容に応じてグループ化する傾向がある

実験結果・課題

■ データセット
　◇ 9ブログ206ページ（コメント率38%）
■ ポストの抽出性能
　◇ 適合率：88.9%　再現率：86.7%　F値：87.7%（+1.5）
　◇ 安定して高い抽出性能
■ コメントの抽出性能
　◇ 適合率：83.8%　再現率：91.3%　F値：87.4%（+5.2）
　◇ コメントが十分あるブログでは高い性能で安定
■ 今後の課題
　◇ 抽出ルールの自動獲得/ルールの自動選択