教師情報を必要としないWebページ群のコンテンツ自動抽出ツールの提案

DEIM Forum 2009 A8-4

Webコンテンツ自動抽出ツール

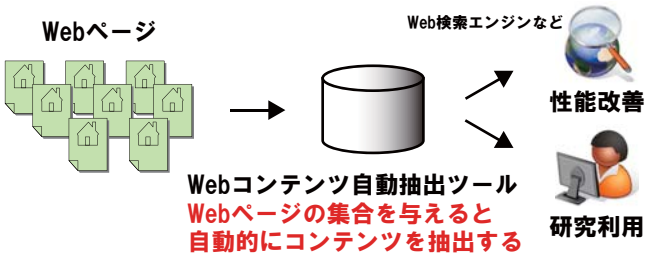
筑波大学
吉田光男　山本幹雄

Webページのコンテンツを利用する研究のお手伝いをします

□ 研究の背景

■ CMS(Contents Management System)の普及
- □ Webページが急増するが不要部分も急増

■ Webページのコンテンツを自動抽出する事で
- □ Web検索エンジン等の性能改善
- □ Webの研究利用を促進



□ 公開物の紹介

■ Webコンテンツ自動抽出ツール（研究成果）

手軽に使えるWebコンテンツ抽出ツール

参考：実行例 $ extract -r /html_dir -P /contents_dir

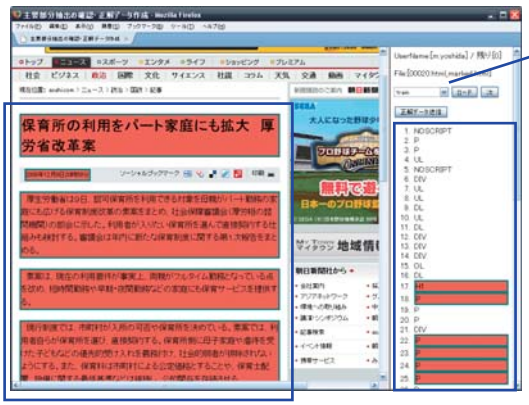
- 学習データの**必要なし**
  Webページ群を対象とする事で実現
- 閾値等のチューニングの**必要なし**
  コンテンツは他のページに出現しないというアイデア
- サイトごとに分ける**必要なし**
  サイトを混合しても性能が変化せず（実験結果より）
- 言語（文字コード）の**影響を受けない**
  特徴ベクトル生成に単語を使わない事で実現

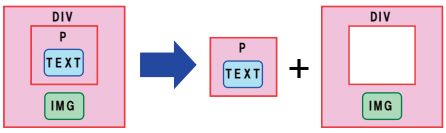
コンテンツ抽出例

■ アノテーションツール（研究補助ツール）

マウス操作のみで行えるラベル付けツール

参考：ラベル付け所要時間 13秒/ページ

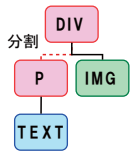




ブロックレベル要素の重なりは分割して認識 ⇔ コンテンツの最小単位を自動的に認識

□ 自動抽出ツールの裏側

アルゴリズム



下位にブロックレベル要素が存在しないようにブロックを抽出

実際には左の例よりも多数のブロックが抽出される

1．ブロックレベル要素を基にブロック抽出

$$Sim(\mathbf{B}_{ij},\mathbf{B}_{kl}) = \frac{\mathbf{B}_{ij}\cdot\mathbf{B}_{kl}}{\|\mathbf{B}_{ij}\|\|\mathbf{B}_{kl}\|}$$

ブロック間の比較はコサイン類似度を用いる
特徴ベクトルは各テキストと各タグの数

コサイン類似度を用いる事で、レンダリングに影響を与えない若干の違いを吸収

2．他のページに出現しないブロックを抽出

実験・課題

■ 実験
- □ データセット：各サイト約180ページ

国内のニュースサイトに対する実験結果(%)

サイト名	適合率	再現率	F値	完全一致率
asahi.com	99.8	97.8	98.8	89.4
毎日.jp	93.7	79.3	85.9	51.1
YOMIURI ONLINE	99.7	95.6	97.6	81.3
合計	98.0	91.1	94.4	73.8

データセット(Webサイト)を混合させた実験結果(%)

サイト名	適合率	再現率	F値	完全一致率
jpALL	98.0	91.1	94.5	73.8

ほぼ一致

- 適合率・再現率が約95%という**高い抽出性能**
- サイトを混合しても**抽出性能に影響を与えない**

■ 今後の課題
- □ 重複Webページの自動検知
  毎日jp 再現率：79% → 98% （人手で重複除去）
- □ 日付表現等の抽出
  バリエーションが少ない情報をどう抽出するか