アンカー関連テキストを用いたWebページ分類

学籍番号：90173024　　西田研究室　　大坪 正典

1　はじめに

2005年現在，Googleが持つデータベースには80億ものWebページがある．膨大な情報の中からユーザが要求する情報を見つけるため，Yahoo!やExciteなどのようなWebページをカテゴリ分類しているポータルサイトの需要が高まっている．しかしこれらのポータルサイトのカテゴリ分類は人手でなされているため，80億ものページを分類することはできない．そこで，Webページを解析し自動で分類しようとする研究が行われてきた．

従来の方法では，分類対象のページ(以下，ターゲットページ)を分析してWebページを自動分類していた．しかしながらターゲットページ中にはそのページを説明するような情報は少ない．そこで，ターゲットページを用いるのではなく，そのページにリンクしているページ(以下，リンク元ページ)を用いて分類する方法が近年注目されてきた．Gloverらはリンク元ページのアンカー前後25単語を用いてWebページ分類を行っている．

しかしながら，リンク元ページを用いたこれまでの研究は，フォーマットに関わらず常に同じルールでアンカー周辺テキストを抽出している．これは，リンク元ページのフォーマットが均一であることを前提としている．そのため，BBSなど様々なスタイルで書かれたWebページ全体に適用するには無理がある．その改善方法として，Webページの DOM構造を見ることが挙げられる．そこで本研究では，DOMレベルの文書構造からアンカーに関連するテキスト部分(以下，アンカー関連テキスト)を推定することを提案する．本研究の目的は，アンカー関連テキストを分類に用いることで，より精度の高い自動分類を目指すことにある．

2　アンカー関連テキスト

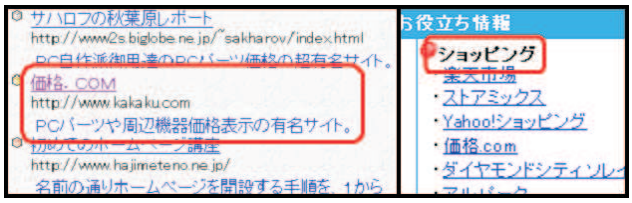

図 1: LSP と USP の例

アンカー関連テキストの抽出方法を決めるにあたり，Hungらの調査結果を用いた．彼らはアンカー関連テキストを大きく”Local Semantic Portion(以下，LSP)”と”Upper-level Semantic Portion(以下，USP)”の2種類に分けた．図1にLSP(左)とUSP(右)の例を示す．いずれもアンカーテキストは”価格.com”である．LSPはアンカーテキストを含むアンカー周辺テキスト部分であり，文書構造上アンカーノードと同レベルにあるものを指す．また，USPはアンカーに接していないテキスト部分で，文書構造上アンカーノードよりも上位レベルに位置するものを指す．Hungらの調査は，LSPについて333ページ，USPについて100ページの計433ページに対して行われ，この調査に基づいてアンカー関連テキストの抽出方法を決めた．以下に，アンカー関連テキストの抽出方法を挙げる．

Local Semantic Portion の抽出方法

- アンカーが段落内，ブロック内にあるとき
  段落中に改行がなければ段落全体を抽出し，改行があれば形式に応じてアンカーを含むテキスト部分を抽出する．
- アンカーがリスト内にあるとき
  アンカーを含む項目全体を抽出する．
- アンカーがテーブル内にあるとき
  アンカーを含むセルの両隣を調べていき，他のアンカーを見つけるまで拡張していく．他のアンカーが見つかったら，その間のセルのテキストを抽出する．

Upper-level Semantic Portion の抽出方法

- タイトル，ヘッダーの抽出
  ページタイトルと，アンカーよりも前にあるヘッダーを抽出する．ただし，同じヘッダーがある場合は最も近くにあるヘッダーを抽出する．
- テーブルヘッダーの抽出
  アンカーがテーブル内にあるとき，テーブルヘッダーがあればテーブルヘッダーを抽出する．テーブルヘッダーがない場合は形式に応じてテーブルの1段目を抽出する．

3　評価用システム

アンカー関連テキストの有用性を確かめるため，Webページの分類精度を求めるシステムをJavaにより実装した．本システムの流れを図2に示す．

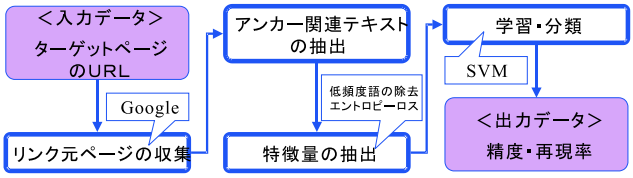

図 2: システムの流れ

ターゲットページのURLをシステムに渡すと，Googleのリンク検索機能を用いてリンク元ページを収集する．次に，リンク元ページよりアンカー関連テキストを抽出し，低頻度語を除去して単語ベクトルを得る．各単語ベクトルのエントロピーロス[1]を計算し，エントロピーロスの大きいものを特徴量として抽出する．最後に特徴量をSVMに学習させ，テスト用データを分類し，精度・再現率を算出する．

4　おわりに

本研究では，Webページの自動分類精度を上げることを目的とし，DOM構造を考慮して抽出したアンカー関連テキストを用いてWebページを分類する手法を提案した．アンカー関連テキストの抽出方法は，Hungらの調査結果に基づいている．また，アンカー関連テキストから特徴量を抽出し，SVMを用いてWebページ分類を行うシステムを実装した．今後は，実装したシステムを用いて評価実験を行い，本手法の有用性を検証する．

[1] (事前エントロピー) − (事後エントロピー) で求められる値．正データ・負データの両方に頻出する単語はエントロピーロスが小さくなる．つまり，エントロピーロスの大きいものを特徴量とすることでSVMの分類精度は上がる．