1G2-1

アンカー関連テキストを用いたWebページ分類方式の実装と評価

Implementing and Evaluating Web Page Classification Method by Anchor-related Text

土方 嘉徳 Yoshinori HIJIKATA　大坪 正典 Masanori OTSUBO　Bui Quang Hung Bui QUANG HUNG　西田 正吾 Shogo NISHIDA

大阪大学大学院 基礎工学研究科
Graduate School of Engineering Science, Osaka University

With the exponential growth of information on the Internet, we need to categorize web pages automatically. Many studies extract keywords from web pages to classify them by using the keywords. Recently, they extract keywords not only from a target page which they want to categorize, but also from the pages which link to the target page. However these approaches conduct the same extraction method even if the format of web pages differs. In our research, we change extraction method by the format of web pages in order to adapt each web page.

1. はじめに

2007 年現在，Google が持つインデックスは 80 億ページ以上といわれている．Yahoo!や Excite に代表されるポータルサイトでは，Web ページを人手でカテゴリ分類し，Web ディレクトリサービスとして提供している．しかし人手では 80 億もの Web ページを処理できないため，Web ページの自動分類に対するニーズが出てきた．

近年の自動分類研究では，分類対象のページ (ターゲットページ) ではなく，そのページにリンクしているページ (リンク元ページ) を利用する手法が注目されている．例えば Glover ら [1] は，リンク元ページのアンカー前後 25 単語を用いて分類を行っている．

しかしこれらの従来研究は，アンカー周辺の形式 (リスト，テーブル etc.) に関わらず，一定部分を抽出している．そこで本研究は，形式に応じて抽出方法を変え，より意味のあるテキスト部分 (アンカー関連テキスト) の抽出を試みる．アンカー関連テキストを分類に用いることで，より正確な自動分類が実現されると考えられる．

2. アンカー関連テキスト

アンカー関連テキストの抽出方式を決めるにあたり，事前調査 [2] を行った．リンク元ページを 1108 ページ収集し，各々のページ中でターゲットページに関する記述部分を人手で調査した．この調査により，アンカー関連テキストには “Local Semantic Portion (LSP)” と “Upper-level Semantic Portion (USP)” の大きく 2 種類があると分かった．図 1 に LSP と USP の例を示す．

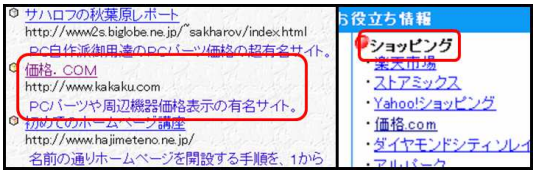

図 1: LSP(左) と USP(右) の例

いずれもアンカーテキストは “価格.com” である．LSP はアンカーを含む周辺テキストであり，文書構造上アンカーと同レベルにあるテキスト部分を指す．また，USP はアンカーと接しておらず，文書構造上アンカーよりも上位にあるテキスト部分を指す．以下，調査結果から決定した LSP と USP の抽出方法を述べる．

2.001　Local Semantic Portion (LSP) の抽出方法

- アンカーが段落 (P タグ) 内にあるとき
  P タグ内に改行がなければ段落全体を抽出．改行があればアンカーを含む部分を抽出．
- アンカーがリスト (OL,UL,DL タグ) 内にあるとき
  アンカーを含む項目 (LI タグ) 全体を抽出．
- アンカーがテーブル (TABLE タグ) 内にあるとき
  アンカーを含むセルの両隣を，他のアンカーが見つかるまで拡張し，その間のセルのテキストを抽出．

2.002　Upper-level Semantic Portion (USP) の抽出方法

- タイトルとアンカーよりも前にあるヘッダーを抽出．(但し，同じヘッダーがある場合は最も近いもの)
- アンカーがテーブル内にあるとき，TH タグがあれば抽出．ない場合はテーブルの 1 段目を抽出する．
- アンカーがリスト内にあるとき，リスト直前のテキスト部分が 20 単語以内であれば抽出．

3. 評価用システム

アンカー関連テキストを Web ページ分類に使用した場合の有用性を確かめるため，Web ページ分類システムを Java で実装した．システムの流れを図 2 に示す．

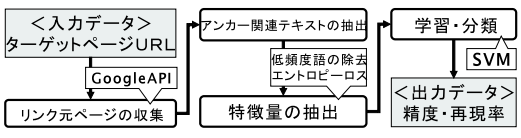

図 2: Web ページ分類システムの流れ

ターゲットページの URL が入力されると，GoogleAPI[3] でリンク元ページを収集し，そこからアンカー関連テキストを

連絡先: 土方嘉徳（大阪大学大学院基礎工学研究科），hijikata@sys.es.osaka-u.ac.jp

抽出する．次に，アンカー関連テキスト中の単語を数え低頻度語を除去し，残った単語のエントロピーロスを計算して，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 (分類の両側に頻出する単語はエントロピーロスが小さくなる)．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. 評価実験

実験に用いるターゲットページを Yahoo! Directory[4] から収集した．負データは全体から 4942 ページ，正データは下記カテゴリからカッコ内のページ数を集めた．

データ 1 Science/Biology (634)

データ 2 Science/Biology/Zoology/Animals (1594)

データ 3 Entertainment/Music (757)

データ 4 Entertainment/Music/Genres/Rock&Pop (634)

データ 5 Recreation/Sports (1090)

データ 6 Recreation/Sports/Baseball/MajorLeague (713)

データ 7 Computers/Internet (953)

データ 8 Computers/Internet/WWW/Weblogs (698)

上記の各データに対して，400 ページ (学習用 300 ページ/テスト用 100 ページ) をランダムに選択し，分類実験を行った．その実験結果を図 3 に示す．ここで，縦軸は F 値 (= (2 ∗ 精度 ∗ 再現率)/(精度 + 再現率)) である．また，各分類手法に用いたテキスト部分を以下に示す．

AT アンカーテキストのみ

All リンク元ページ中の全テキスト

TP ターゲットページ中の全テキスト

Fix25 アンカー前後 25 単語

STP アンカー関連テキスト（LSP+USP）

Mix USP+Fix25

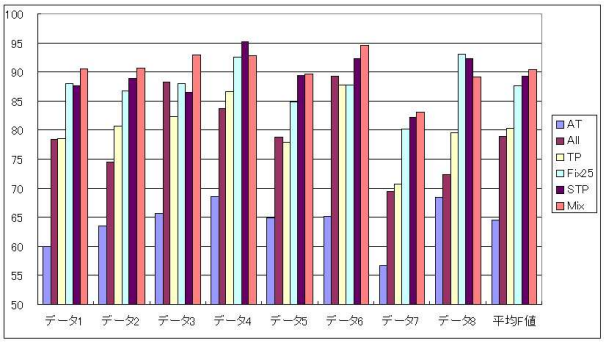

図 3: データ別/手法別 - F 値

提案手法である STP の平均 F 値は 89.2，Glover ら [1] の手法である Fix25 の平均 F 値は 87.6 であった．このことから提案手法であるアンカー関連テキストが Web ページ分類に有効であることが分かった．また，Mix 手法として USP と Fix25 を組み合わせたテキストを用いた場合 90.4 の平均 F 値を得た．

次に，アンカー関連テキストの内，LSP と USP のどちらが Web ページ分類に有効かを調べるため，それぞれ単独で分類実験を行った結果を図 4 に示す．

分類実験の結果，LSP よりも USP の方が分類に貢献することが分かった．しかしながら，USP よりも STP の方が分類精度が良いため，LSP が分類精度を下げている訳ではなく，LSP と USP を組み合わせて用いた方がよいことが分かった．

更に，Glover ら [1] が提案したアンカー前後 25 単語を用いた Web ページ分類について「前後 25 単語」という数字が最適なのかを調べるための実験を行った（図 5）．

結果としては，25 単語が最適であるという訳ではなく，25 単語よりも単語量が多い方が結果が良くなる傾向にあることが分かった．この実験に関しては，より前後単語量を増やして，最適な単語量を調べていく必要がある．

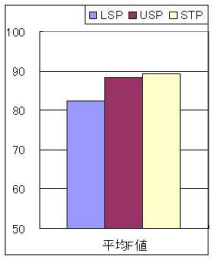

図 4: LSP/USP 別 - 平均 F 値

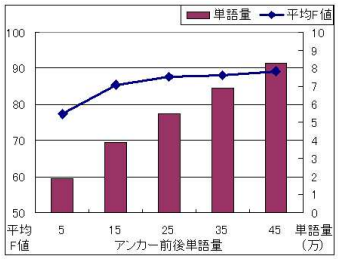

図 5: アンカー前後の単語量を変化 - 平均 F 値/単語量

5. おわりに

本稿では，文書構造を考慮して抽出するアンカー関連テキストを用いて Web ページ分類することを提案した．また 3 つの評価実験を行い，アンカー関連テキストの Web ページ分類に対する有用性や，LSP/USP を単独で用いた場合の分類精度，アンカー前後 25 単語の適性について調べた．

今後は，学習データの正負の比率を変え，より安定した分類を目指す．また，アンカー前後の単語量がどのくらいで最適となるのか調べることも考えられる．更に，今回は英語ページを対象としているが，日本語ページにも対応したい．

参考文献

[1] Eric J.Glover, Kostas Tsioutsiouliklis, Steve Lawrence, David M.Pennock, Gary W.Flake: Using Web Structure for Classifying and Describing Web Pages, WWW2002, pages 562-569, Hawaii, USA, 2002.

[2] Bui Quang Hung, Masanori Otsubo, Yoshinori Hijikata, Shogo Nishida: Extraction of Semantic Text Portion Related to Anchor Link, IEICE Transactions on Information and Systems, pages 1834-1847, VOL.E89D, NO.6 JUNE 2006.

[3] Google SOAP Search API (Beta), http://code.google.com/apis/soapsearch/

[4] Yahoo! Directory, http://dir.yahoo.com/