ハイパーリンクを用いたWeb文書の自動分類

鈴木 祐介† 松原 茂樹‡ 吉川 正俊‡

†名古屋大学大学院情報科学研究科 ‡名古屋大学情報連携基盤センター

suzuki@dl.itc.nagoya-u.ac.jp

1 はじめに

ディレクトリ型検索エンジンは，あらかじめ Web 文書をトピックごとに分類しておくことにより，ユーザの目的にあった文書へのアクセスを容易にする．従来，Web 文書のディレクトリへの分類は人手によるものがほとんどであったが，Web 文書の急速な増加に伴い，分類の自動化が求められている．

文書分類においては，特徴素の抽出方法が分類精度に大きく影響する [1, 2, 3]．これまでに提案された抽出方法の多くは，分類の対象となる文書内に出現する単語に重みを与えることにより，文書を特徴付けている [4, 5]．しかし，Web 文書は文書内容やそのサイズも様々であるため，対象とする Web 文書のテキスト量が少ないときには，分類に有効な特徴素を得ることは難しい．また，Web 文書は画像や映像といったテキストをもたないコンテンツも含んでおり，これらを自動で分類するための手法が求められる．

本稿では，Web 文書の自動分類における分類精度の向上のため，ハイパーリンクによる参照情報を用いた文書分類手法を提案する．Web 文書では，アンカーテキストはリンク先の文書内容を端的に表現していることが多く，そこから得られる特徴素はリンク先の文書分類に有効に機能すると考えられる．そこで本手法では，分類する文書そのものから特徴素を抽出する以外に，分類の対象となる文書の特徴素を，リンクで直接参照している文書群のアンカーテキストとその周辺のテキストからも抽出する．すなわち，分類対象の Web 文書とそのリンク元の Web 文書群から作成する 2 つの特徴ベクトルによって文書を表現する．以下，本稿では，分類する Web 文書を分類対象文書，分類対象文書をリンクで直接参照する Web 文書を参照元文書と呼ぶことにする．

本手法の有効性を評価するために，名古屋大学 Web ディレクトリを用いて分類実験を行った．実験の結果，HTML 文書の分類においては，本手法の正解率は 59.5% であり，従来手法と同程度であったものの，画像文書の分類において 43.5%の正解率を示しており，Web 文書分類における本手法の有効性を確認した．

2 ベクトル空間モデル基づく分類

文書を分類するときの基準となるカテゴリの特徴ベクトルは，あらかじめ分類済みの文書集合から作成する．カテゴリに属する分類済みの文書集合を 1 つの文書とみなして単語を抽出したのち，各カテゴリごとに，TF-IDF 法を用いて単語を重み付けする．カテゴリ $C_i(1 \leq i \leq M)$ における単語 $e_j(j = 1, 2, \cdots, N)$ の TF-IDF 法による重み w_{ij} を式 (1) より求め，カテゴリ C_i の特徴ベクトル $\vec{x_i} = (w_{i1}, w_{i2}, \cdots, w_{iN})$ を作成する．

$$w_{ij} = F_{ij} \times log\frac{M}{v_j} \tag{1}$$

なお，F_{ij} はカテゴリ C_i に出現する単語 e_j の頻度，v_j は単語 e_j を含むカテゴリ数，M は全カテゴリ数を表す．カテゴリの特徴ベクトルはその大きさが 1 になるように式 (2) に従って正規化する．

$$\vec{x_i}' = \frac{\vec{x_i}}{|\vec{x_i}|} \tag{2}$$

次に，文書の特徴ベクトルは分類先が未定の各文書ごとに作成する．分類対象文書から単語を抽出し，その単語の頻度を用いて重み付けする．文書 d における単語 $e_j(j = 1, 2, \cdots, N)$ の重み w_{d_j} を式 (3) により求め，文書の特徴ベクトル $\vec{d} = (w_{d_1}, w_{d_2}, \cdots, w_{d_N})$ を作成する．

$$w_{d_j} = F_{d_j} \tag{3}$$

なお，F_{d_j} は文書 d に出現する単語 e_j の頻度を表す．

文書の分類先は，文書 d のカテゴリ C_i に対する類似度を式 (4) で求め，その値が最大となるカテゴリであるとして定める．

$$Sim(d, C_i) = \vec{x_i}' \bullet \vec{d} \tag{4}$$

3 参照情報を用いたWeb文書の表現

3.1 提案手法の概要

Web 上の文書群は図 1 のようにハイパーリンクで互いに関連付けられているため，ユーザはリンクをたどり

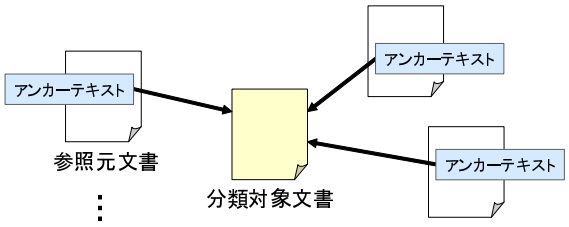

図 1: Web 文書間のハイパーリンクによる関係

ながら目的の文書を見つけ出す．このとき，ユーザは，現在閲覧している文書中の情報をもとにリンク先文書の内容を推測することになる．特に，アンカーテキストはリンク先文書を直接参照する部分でもあるため，リンク先文書の内容を端的に表現していることが多く，その内容を判断する上で有効な情報となる．また，アンカーテキストだけでは内容を判断できない場合でも，ユーザは必要な情報をアンカーテキスト周辺の記述から獲得し，リンク先文書の内容を推測している．

このため，参照元文書からリンク先文書について端的に記述された箇所を特定し，そこから特徴素を抽出すれば，リンク先文書の分類において効果的に機能すると考えられる．また，参照元文書から特徴素を抽出することにより，テキストをもたない画像や映像等の文書の分類も可能になる．

本研究では，従来のように分類対象文書から特徴素を抽出する以外に，分類対象文書を直接リンクで参照している参照元文書群からも抽出する．なお，参照元文書から特徴素を抽出するときには，その文書内で分類対象文書について記述している箇所を特定し，その箇所から抽出する．Web 文書は分類対象文書とその参照元文書群から別々に作成した 2 つの特徴ベクトルで表現する．

本節の残りは，参照元文書から特徴素を抽出する手順を説明したのち，抽出した特徴素を用いて分類対象文書の特徴ベクトルを作成し，分類する方法について述べる．

3.2 参照元文書からの特徴素抽出

参照元文書からの特徴素の抽出には，分類対象文書について記述している箇所のみを利用する．抽出手順としては，まず，HTML タグを手がかりに，特徴素の抽出箇所を特定し，次に，特定した箇所の形態素解析結果をもとに名詞を特徴素として抽出する．以下に，抽出箇所を特定するときに使用する HTML タグについて，HTML テキストのサンプルを用いて説明する．サンプルでは，“file.html”が分類対象文書であるとし，下線部分がアンカーテキスト，網掛け部分が抽出箇所を表す．

1）アンカーテキスト

アンカーテキストはリンク先の文書を直接参照しており，分類対象文書の内容を最も端的に表現した部分である．本手法では，アンカータグ (<A>) で囲まれたテキスト部分を抽出箇所として取り出す．

```
<a href="file.html">後期課程募集要項</a>
```

また，リンクに画像が使用されている場合は，画像の ALT 属性もアンカーテキストとみなし，ALT 属性のテキストも抽出箇所として取り出す．

```
<a href="file.html">
<img src="image.gif" alt="過去の入試問題">
</a>
```

2）タイトル

Web 文書のタイトルはそのページ全体の内容を簡潔に表現していることが多いため，そのページに含まれるアンカーテキストの内容を補足すると考えられる．ここでは，タイトルタグ (<TITLE>) で囲まれたテキスト部分を抽出箇所として取り出す．

```
<title>入試情報</title>
                :
<a href="file.html">後期課程募集要項</a>
```

3）見出し文字

見出し文字はそのページの部分文書における主題を表すために使用されることが多いため，リンクがその部分文書に含まれる場合はアンカーテキストの内容を補足すると考えられる．見出し文字はヘッダタグを用いて表現されることが多く，ヘッダタグは <H1> を最上位レベルとして，以下 <H6> まで数字が大きくなるにつれて文字のサイズが小さい下位レベルになる．そこで，対象とするリンクの位置から遡っていき，対象とするリンクと最も近い位置にあるヘッダタグで囲まれたテキスト部分を抽出箇所として取り出す．続いてさらに溯り，そのヘッダタグよりも上位レベルのヘッダタグが存在していれば，そのヘッダタグで囲まれたテキスト部分も抽出箇所として取り出す．

```
<h1>大学院<h1>
<h2>博士課程 前期課程</h2>
       :
<h2>博士課程 後期課程</h2>
<a href="file.html">後期課程募集要項</a>
```

3.3 Web 文書の表現

本手法では，Web 文書は分類対象文書とその参照元文書群から別々に作成する 2 つの特徴ベクトルを用いて表現する．分類対象文書からの特徴ベクトルは 2 節で求めた文書の特徴ベクトルを用いる．本節では，参照元文書群から特徴ベクトルを作成する方法について述べる．

まず，分類対象文書をリンクで直接参照している Web 文書をすべて集める．集めた Web 文書群から 3.2 節に基づいて単語を抽出し，その単語の頻度に抽出箇所による重みを掛けることにより参照元文書群の特徴ベクトルを作成する．参照元文書群 r における単語 $e_j(j = 1, 2, \cdots, N)$ の重み w_{r_j} を式 (5) で定義すると，参照元文書群による Web 文書の特徴ベクトルは，

$$\vec{r} = (w_{r_1}, w_{r_2}, \cdots, w_{r_N})$$

で表される．

$$w_{r_j} = \sum_t (F^t_{r_j} \times k_t) \tag{5}$$

ここで，$F^t_{r_j}$ は参照元文書群 r の抽出箇所 t に出現する単語 e_j の頻度，k_t は抽出箇所に基づく単語の重み値を表す．なお，抽出箇所 t は，アンカーテキスト，タイトル，見出し文字のことである．

Web 文書は 2 つの特徴ベクトル $\vec{d}$ と $\vec{r}$ で表現する．2 つの特徴ベクトルはベクトルの大きさが 1 になるように，式 (6) に従って正規化する．

$$\vec{d'} = \frac{\vec{d}}{|\vec{d}|} \qquad \vec{r'} = \frac{\vec{r}}{|\vec{r}|} \tag{6}$$

3.4 分類方法

Web 文書 d のカテゴリ C_i における類似度は，分類対象文書と参照元文書群のそれぞれの特徴ベクトルと 2 節で求めたカテゴリの特徴ベクトルによる内積を計算したのち，式 (7) により求め，その値が最大となるカテゴリに分類対象文書を分類する．なお，α はパラメタである $(0 \leq \alpha \leq 1)$.

$$Sim_{Doc}(d, C_i) = \vec{x_i}' \bullet \vec{d'}$$
$$Sim_{Ref}(d, C_i) = \vec{x_i}' \bullet \vec{r'}$$
$$Sim(d, C_i) = (1 - \alpha) Sim_{Doc}(d, C_i) + \alpha Sim_{Ref}(d, C_i) \tag{7}$$

4 評価実験

4.1 実験の概要

4.1.1 実験データ

本実験では，分類対象として名古屋大学 Web ディレクトリを使用する．このディレクトリは名古屋大学ドメインの Web 文書が内容に応じてカテゴリごとに分類されている．カテゴリは階層的に構成されており，階層が深くなるにつれて内容はより詳細になる．例えば，大学院前期課程の募集要項に関する文書は「大学に入る - 大学院 (前期課程) - 募集要項」に属する．なお，カテゴリは 9 つのトップカテゴリからなり，深さは最大で 3 階層である．名古屋大学 Web ディレクトリの概観を図 2 に示す．

図 2: 名古屋大学 Web ディレクトリ

本実験では，このディレクトリに分類されている日本語で記述された Web 文書 5684 ページを使用し，分類先カテゴリは実験に使用する Web 文書が実際に分類されている 126 カテゴリを対象とした．使用する Web 文書は平均 1.07 個のカテゴリに分類されている．本実験では，Web 文書をカテゴリの特徴ベクトル作成用に 5169 ページ，評価用に 515 ページに分割して使用した．なお，評価用ページのファイルタイプは HTML 文書 430 ページ，画像ファイル 85 ページからなる．

4.1.2 参照元文書の獲得

参照元文書から特徴素を抽出するために，分類対象文書における参照元文書を Web から収集した．本実験では，収集する参照元文書の範囲を名古屋大学ドメインの Web 文書とし，ドメイン内の Web 文書をすべて集めてリンク解析を行うことにより参照元文書を割り出した．

4.1.3 評価手法

評価は，本手法を分類対象文書のみから特徴素を抽出する従来手法と比較することにより行う．ただし，正解の分類先が 2 つ以上ある場合はそのうちのいずれかに分類されれば正解であるとした．また，分類対象文書の特徴素はタイトルタグ (<TITLE>)，ヘッダタグ (<H1>～<H6>)，強調タグ (<B>) で囲まれたテキスト部分と BODY タグ (<BODY>) で囲まれたテキストの開始

表 1: 抽出箇所による単語の重み

分類対象文書		参照元文書	
抽出箇所	重み	抽出箇所	重み
タイトル	1	アンカーテキスト	2
見出し文字	1	タイトル	1
強調文字	1	見出し文字	1
本文 (300)	1		

300 文字より抽出した．抽出箇所による単語の重みは表 1 のように設定し，式 (3)，及び，式 (5) による単語の頻度に対して重みを与えた．また，式 (7) のパラメタ α は 0.2 とした．なお，実験では，形態素解析器に「茶筌」[6] を使用している．

4.2 結果と考察

実験結果を表 2 に示す．HTML 文書を分類した場合，本手法による正解率は 59.5%となり，従来手法と同程度の正解率であった．しかし，本手法では，従来手法では分類できない画像ファイルも分類しており，その正解率は 43.5%であった．

本手法の結果は，テキストをもたない Web 文書の分類可能性を示唆しているものの，HTML 文書の分類においては，参照元文書の特徴素を追加することの効果があまり現れていない．以下，本実験を通して観察された参照元文書の情報を用いることの問題点について考察する．

参照元文書からの特徴素の抽出は，分類対象文書から得られる特徴素が少ない場合でも，他の文書から広く特徴素を抽出できる利点がある．しかし，多くの Web 文書から参照されるドメインのトップページに比べ，内部的なページではその参照元ページは同一ドメイン内のページに限られ，その数も極端に少なくなる．また，リンクに画像を使用しているが，ALT 属性を設定していないサイトも見受けられる．このような場合，参照元文書から十分な情報を得ることは難しくなる．

アンカーテキストの記述には,「戻る」や「Top へ」といった案内語や「こちら」といった指示語も見られる．前者はリンク先文書の内容を表現しておらず，文書間の関連性も薄くなる．また，後者の記述からは分類に有効な特徴素を得ることはできないものの，アンカーテキスト前後の文脈にはリンク先文書に関する内容が記述されていることが多い．実験では，上記以外にもアンカーテキストに人名や論文名，広報誌等の発行表記 (例えば,「No.139」や「平成 15 年度」) などが使われている場合は，正しく分類されないことが多かった．ただし，参照元文書に論文名などが一覧表示されていれば，本文中の「論文一覧」といった見出しやタイトルを抽出することにより，正しく分類されるようになった例もいくつか観察された．このことから，参照元文書を用いて分類精度を向上させるためには，アンカーテキストの分類への効果を見極めるとともに，分類対象文書についての記述をアンカーテキスト以外からも的確に捉えることが重要になると言える．

表 2: 従来手法との比較

手法	HTML		画像	
	正解数	正解率 (%)	正解数	正解率 (%)
提案手法	256	59.5	37	43.5
従来手法	257	59.8	—	—

5 まとめ

本稿では，ハイパーリンクによる参照情報を用いた Web 文書分類手法を提案した．文書分類に使用する特徴素を参照元文書のアンカーテキストとその周辺テキストからも抽出することにより，テキストを含まない文書ファイルの分類も可能になった．

今後の課題としては，文書分類への効果の観点から，アンカーテキストのタイプを分析することが重要になる．また，分類正解率を上げるために階層カテゴリの構造を考慮し，カテゴリ間の関係を用いた分類手法を検討する必要がある．

参考文献

[1] 石田 栄美: テキストの自動分類の要素分析的アプローチ, 情報処理学会研究報告，FI65-4，pp.33-40，(2001).

[2] John M. Pierre: Practical Issues for Automated Categorization of Web Sites , In *Proceedings of the ECDL 2000 Workshop on the Semantic Web* (2000).

[3] 森本 由起子，間瀬 久雄，辻 洋: 記事データからの分類知識獲得に関する実験シミュレーション, 情報処理学会研究報告，DD6-1，pp.1-8 (1997).

[4] 安形 輝，石田 栄美，久野 高志，野末 道子，上田 修一: WWW ページの自動分類　NDC の分類体系と Yahoo のカテゴリを使った分類，情報処理学会研究報告，FI54-15, pp.113-120 (1999).

[5] 福本 文代，鈴木 良弥: 語の重み付け学習を用いた文書の自動分類, 情報処理学会論文誌, Vol.40, No.4, pp.1782-1791 (1999).

[6] 松本 裕治，北内 啓，山下 達雄，平野 善隆，松田 寛，高岡 一馬，浅原 正幸: 形態素解析システム『茶筌』, version2.2.9, 使用説明書 (2002).