ディレクトリ型検索エンジンにおける Webページのカテゴリ自動分類

細川 誠介　　的場 隆一
富山高等専門学校 電子情報工学科
{i1211431, rmatoba}@nc-toyama.ac.jp

1 はじめに

我々はWebページから目的の情報を得る際，Google等のロボット型検索エンジンやYahoo!カテゴリ等のディレクトリ型検索エンジンを使用してWebページの検索を行うことが多い．

ロボット型検索エンジンは，ロボットがWebページを自動で巡回して情報を収集するので，膨大な範囲の情報から検索することができる．しかし，検索語を一つでも含むWebページのリストをジャンルによらず表示するので，ユーザーにとって不適切なWebページを表示する可能性があるという欠点がある．

また，ディレクトリ型検索エンジンは，Webページが項目別にまとめられているので，同じジャンルのWebページや関連したジャンルのWebページを容易に検索することができる．しかし，Webページの分類に人員やコストがかかり，登録されているWebページの数が少ないという欠点がある．

これらの検索エンジンの欠点を補うためにWebページをカテゴリへ自動分類する手法が存在する [1][2][3]．佐々木らの研究 [4] では，スポーツ以下のサブカテゴリを対象としてmeta要素の内容を取得し，ナイーブベイズ分類により82%の分類精度が得られたと報告されている．

しかし，実際にはmeta要素の無いWebページが多く存在する．例えば，Yahoo!カテゴリに登録されているWebページのうち49.8%のWebページにはmeta要素が記載されていなかった．また，実用性を考慮すると，任意の分野のWebページをカテゴリに分類できるシステムを構築するよりも，あらゆる分野のWebページをカテゴリに分類できるシステムを構築することが望ましいといえる．

そこで，本研究では，Yahoo! カテゴリに存在する13個のトップカテゴリを対象として，meta要素を用いずにtitle要素とbody要素の内容を取得し，ナイーブベイズ分類を用いて実用的なWebページのカテゴリへの自動分類を行うシステムを提案する．

2 ナイーブベイズ分類

ナイーブベイズ分類では，文書 D が与えられた時，カテゴリ C に属する確率 $P(C|D)$ を式 (1) で表す．

$$P(C|D) = \frac{P(C)P(D|C)}{P(D)} \tag{1}$$

このとき，文書 D を bag-of-words で単語の集合 $W = \{W_1, W_2, \cdots, W_n\}$ として考えると，$P(D|C)$ は式 (2) で表される．このとき，単語 $W_1, W_2, \cdots, W_n$ はそれぞれ独立して出現するものと仮定する．

$$\begin{aligned} P(D|C) &= P(W_1, W_2, \cdots, W_n|C) \\ &= \prod_{i=0}^{n} P(W_i|C) \end{aligned} \tag{2}$$

よって，$P(C|D)$ は式 (3) で表される．

$$P(C|D) = \frac{P(C)\prod_{i=0}^{n} P(W_i|C)}{P(D)} \tag{3}$$

式 (3) において，$P(D)$ は文書が生起する確率を表しカテゴリによらず一定の値をとるため，式 (4) で表される $score$ が最大のカテゴリが推定カテゴリとなる．

$$score = P(C)\prod_{i=0}^{n} P(W_i|C) \tag{4}$$

ここで，カテゴリ C に現れる単語 W_i の回数を $T(C, W_i)$，カテゴリ C に含まれる総単語数を $\sum_{W'} T(C, W')$ とすると，$P(W_i|C)$ は式 (5) で表される．

$$P(W_i|C) = \frac{T(C, W_i)}{\sum_{W'} T(C, W')} \tag{5}$$

ただし，この式を使用すると，データベースにない単語が出現した場合に $P(W_i|C) = 0$ となってしまうゼロ頻度問題が起きてしまう．これを避けるためにラプラススムージングを用いて式 (6) で $P(W_i|C)$ を計算する．ただし，V は全てのカテゴリの文書に含まれる異なる単語の数とする．

$$P(W_i|C) = \frac{T(C, W_i) + 1}{\sum_{W'} T(C, W') + V} \tag{6}$$

また，*score* の計算において，$\prod_{i=0}^{n} P(W_i|C)$ は非常に小さな値であるため，アンダーフローを起こす可能性がある．そのため，$P(C)$ と $P(W_i|C)$ の対数をとって総乗の式を総和の式に変換することで *score* を式 (7) のように計算する．

$$score = \log P(C) + \sum_{i=0}^{n} logP(W_i|C) \tag{7}$$

3 自動分類システム

3.1 概要

本研究では，Yahoo!カテゴリに登録されている Web ページをダウンロードして作成したデータベースを基に，Web ページのトップカテゴリを推定する．本研究で提案する自動分類システムの流れを図 1 に示す．

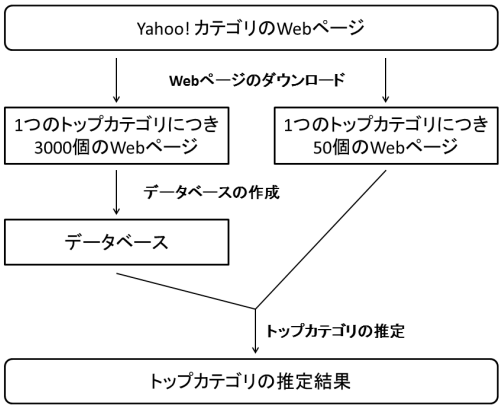

図 1: 自動分類システムの流れ

本節では，図 1 に示した自動分類システムについて，具体的な手順を示す．

3.2 Web ページのダウンロード

本研究では，Yahoo! カテゴリに登録されている Web ページをそれぞれのトップカテゴリにつき 3050 個ダウンロードする．このうち，3000 個の Web ページをデータベースの作成に使用し，50 個の Web ページをカテゴリの推定に使用する．ただし，地域情報のトップカテゴリについては，地域別にエンターテインメント, メディアとニュースなどのカテゴリに分かれていて分類が困難であるため，本研究の対象外とした．

Yahoo! カテゴリに存在するトップカテゴリのうち，地域情報のトップカテゴリを除いた 13 個のトップカテゴリに含まれる Web ページの数を表 1 に示す．

表 1: トップカテゴリに含まれる Web ページの数

トップカテゴリの名前	Web ページの数
エンターテインメント	55294
メディアとニュース	11818
趣味とスポーツ	37519
ビジネスと経済	76609
各種資料と情報源	1177
生活と文化	15793
芸術と人文	14578
コンピュータとインターネット	22627
健康と医学	5339
教育	2942
政治	4812
自然科学と技術	7254
社会科学	2682

表 1 のうち Web ページの数が 3050 個に満たないカテゴリについては，それぞれのトップカテゴリに含まれる Web ページの頻出単語で Google 検索を行い，不足分については手作業で Web ページの分類を行った．また，実際には 12.5%の Web ページが複数のトップカテゴリに属しているが，本研究では全ての Web ページが単一のトップカテゴリにのみ属しているものとしてトップカテゴリの推定を行う．

3.3 データベースの作成

Yahoo! カテゴリに登録されている Web ページをそれぞれのトップカテゴリにつき 3050 個ダウンロードした後，それぞれのトップカテゴリで任意の単語数に達するまで下記に示す 3 つのデータベースを作成する．

- 任意のトップカテゴリに含まれる単語の数をカウントして数値を保存するデータベース

- 任意のトップカテゴリに含まれる Web ページの数をカウントして数値を保存するデータベース
- データベースの作成に使用した全ての単語から重複を除去して単語を保存するデータベース

これらのデータベースを作成するために，ダウンロードした Web ページの中からランダムに Web ページを 3000 個選んで使用する．このとき，HTML の title 要素と body 要素に含まれる任意の種類の単語を形態素解析エンジン MeCab を用いて抽出する．ただし，body 要素の内容のうち，CSS や JavaScript を記述するための要素である style 要素と script 要素の内容は抽出しない．

3.4 トップカテゴリの推定

データベースの作成が終了した後，前述したデータベースを読み込んだうえで，前節で述べたナイーブベイズ分類を用いて Web ページのトップカテゴリを推定する．このとき，ダウンロードした Web ページのうちデータベースの作成に使用していない 50 個の Web ページについてトップカテゴリの推定を行う．

これらの Web ページについては，辞書の作成と同様に HTML の title 要素と body 要素に含まれる任意の種類の単語を抽出して，トップカテゴリの推定に使用する．

4 自動分類実験

4.1 実験設定

3 節で述べた自動分類システムを構築し，2 つの実験を行った．

まず，抽出する品詞の種類およびデータベースの作成に使用する単語数を変化させてそれぞれの条件下での正解率を測定する実験を行った．この実験は，抽出する品詞の種類と正解率の関係やデータベースの作成に使用する単語数と正解率の関係を調べることを目的としている．

また，抽出する品詞の種類およびデータベースの作成に使用する単語数を変化させてそれぞれの条件下でのトップカテゴリ別の正解率を測定する実験を行った．この実験は，トップカテゴリと正解率の関係を調べることを目的としている．

ただし，正解率はそれぞれの条件下でデータベースの作成とトップカテゴリの推定を 10 回試行し，それぞれの試行の正解率を平均した値とした．

4.2 品詞とデータベースの単語数に対する実験結果

抽出する品詞の種類およびデータベースの作成に使用する単語数に対する正解率を表 2 に示す．

表 2: 品詞とデータベースの単語数に対する正解率

品詞	データベースの単語数	正解率
全ての品詞	50 万語	62.7%
	100 万語	65.2%
	150 万語	66.1%
	200 万語	66.9%
名詞	20 万語	64.6%
	40 万語	67.0%
	60 万語	68.4%
	80 万語	69.0%
	100 万語	69.3%
固有名詞 一般名詞	10 万語	65.5%
	20 万語	68.4%
	30 万語	69.3%
	40 万語	69.8%
	50 万語	70.4%
固有名詞	2 万語	37.4%
	4 万語	40.8%
	6 万語	43.6%
	8 万語	44.2%
	10 万語	44.8%
一般名詞	10 万語	68.6%
	20 万語	69.6%
	30 万語	70.7%
	40 万語	71.2%
動詞	2 万語	34.2%
	4 万語	34.8%
	6 万語	35.9%
	8 万語	36.4%
	10 万語	36.5%
形容詞	1000 語	16.6%
	3000 語	17.6%
	4500 語	18.1%
	6000 語	18.7%

表 2 より，一般名詞を 40 万語登録したデータベースを使用した場合に最も正解率が高くなることや，データベースの作成に使用する単語数を増加させることによって正解率が上昇することが確認された．

また，抽出する品詞の種類に対する正解率を比較す

ると，全ての品詞，名詞，固有名詞と一般名詞，一般名詞を抽出した場合に正解率が 60%以上の高い確率になり，固有名詞，動詞，形容詞を抽出した場合に正解率が 50%未満の低い確率になった．

4.3 トップカテゴリ別の実験結果

前述した品詞とデータベースの単語数に対する実験結果より，一般名詞を 40 万語登録したデータベースを使用した場合に最も正解率が高くなることが確認された．この場合のトップカテゴリ別の正解率を表 3 に示す．

表 3: トップカテゴリ別の正解率

トップカテゴリの名前	正解率
エンターテインメント	75.6%
メディアとニュース	76.0%
趣味とスポーツ	70.6%
ビジネスと経済	79.8%
各種資料と情報源	57.2%
生活と文化	64.0%
芸術と人文	83.8%
コンピュータとインターネット	85.6%
健康と医学	77.6%
教育	68.2%
政治	74.8%
自然科学と技術	59.8%
社会科学	52.0%

表 3 より，コンピュータとインターネットのトップカテゴリで最も正解率が高くなることや，いずれのトップカテゴリにおいても正解率が 50%以上となることが確認された．

また，トップカテゴリによって正解率に大きな差があり，芸術と人文，コンピュータとインターネットの 2 つのトップカテゴリでは正解率が 80%以上となり，各種資料と情報源，自然科学と技術，社会科学の 3 つのトップカテゴリでは正解率が 60%未満となった．

5 考察

5.1 品詞とデータベースの単語数の考察

表 2 より，抽出する単語の種類によらず，データベースの作成に使用する単語数を増加させることによって正解率が上昇するものの，その上昇幅は少しずつ小さくなる傾向が見られた．また，一般名詞を抽出した場合に最も正解率が高くなり，一般名詞は各トップカテゴリの特徴を最も表しているといえる．一方で，固有名詞，動詞，形容詞を抽出した場合には正解率が低くなった．これは，これらの品詞が各トップカテゴリの特徴をほとんど表していないことや Web ページに含まれる単語数が少ないことが原因と考えられる．

5.2 トップカテゴリ別の考察

表 3 より，多くのトップカテゴリで正解率が 70%以上となったが，各種資料と情報源，自然科学と技術，社会科学の 3 つのトップカテゴリでは正解率が 60%未満となった．例として，自然科学と技術のトップカテゴリの Web ページ場合は教育のトップカテゴリに分類されることが多かった．これは，科学に関する教育機関の Web ページが教育のトップカテゴリに分類される傾向があるからであると考えられる．

6 おわりに

本研究では，ナイーブベイズ分類を用いた Web ページのカテゴリへの自動分類を行うシステムについて提案した．その結果，71.2%の確率で Web ページを正しいカテゴリに自動で分類ことができた．今後は，1 つの Web ページに対して複数のカテゴリを推定結果として表示することで，複数のカテゴリに属している Web ページの分類にも対応し，より実用的なシステムの構築を目指していきたい．

参考文献

[1] 塚田 誠, 鷲尾 隆, 元田 浩, “機械学習による Web ページの自動分類”, 電子情報通信学会技術研究報告 人工知能と知識処理, pp63-70, 2000

[2] 金村 和美, 力宗 幸男, “Web ページの自動分類に関する一手法”, 電子情報通信学会技術研究報告 オフィスインフォメーションシステム, pp25-30, 2004

[3] 佐々木 稔, 新納 浩幸, “Web サイトの階層的な Web ディレクトリへの自動分類手法”, 情報処理学会研究報告自然言語処理, pp109-114, 2007

[4] 佐々木 稔, 新納 浩幸, “meta タグを利用した Web ディレクトリの自動構築手法”, 言語処理学会第 13 回年次大会論文集, pp895-898, 2007