社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2005－ＩＣＳ－140（1）
2005／8／1

Webサイトの自動分類に向けた特徴分析とキーワード抽出に関する研究

本田　崇智†　　山本　雅人†　　川村　秀憲†　　大内　東†

† 北海道大学大学院情報科学研究科
〒 060-0814 北海道札幌市北区北 14 条西 9 丁目
E-mail: †{honda,masahito,kawamura,ohuchi}@complex.eng.hokudai.ac.jp

あらまし　Web サイトは，ディレクトリ型検索エンジンでも見られるように様々なカテゴリーに属している．一方 Web サイトには，画像ファイルやリンク，テキストデータなどといった多くの情報が存在する．このような情報から得られる特徴やサイト中に含まれるキーワードによって，Web サイトはカテゴリーへと自動分類することが可能であると考えられる．そこで本研究では，宿泊施設や飲食店など観光に関するカテゴリーを対象とし，各カテゴリーごとに特異的に見られる Web サイトの特徴やキーワードの抽出を行い，それらを用いて自動分類の評価を行う．本手法が確立されれば，任意の Web サイトをカテゴリーに自動分類することが可能になると考えられる．

キーワード　Web サイト，自動分類，特徴分析，キーワード抽出

Feature Analysis and Keyword Extraction for Automatic Classification of Web Site

Takatomo HONDA†, Masahito YAMAMOTO†, Hidenori KAWAMURA†, and Azuma OHUCHI†

† Graduate School of Information Science and Technology, Hokkaido University Kita 14, Nishi 9, Kita-ku,Sapporo,060-0814 Japan
E-mail: †{honda,masahito,kawamura,ohuchi}@complex.eng.hokudai.ac.jp

Abstract　A website is considered to be belonging to a certain category such as accommodations, restaurants, facilities and so on, as shown in the directory-typed search engines. Generally, a website includes valuable information such as links to or from other sites, many files with extensions and many text data, in terms of the category classification. In this paper, we investigate whether many websites belonging to a certain category have some common features or not. In particular, we show that some keywords are very important to classify websites to the categories. By using these analyzed keyword information, we present that many websites could be classified to an appropriate category with high precision when three categories (museum, restaurant and accommodations) related to tourism are treated as examples.

Key words　Web site, Automatic Classification, Feature Analysis, Keyword Extraction

1. はじめに

近年，World Wide Web(WWW) の世界は，急激に成長し巨大な情報空間を形成しつつある．そのため，Web を閲覧する場合に求める情報を得ることは容易ではない．そこで検索エンジンと呼ばれるシステムにキーワードを入力することによって，Web の検索を行うのが一般的である．この検索エンジンは大きく分けて，ディレクトリ型検索エンジンとロボット型検索エンジンの 2 つに分けられる．

ディレクトリ型検索エンジンとは，Web サイト群をカテゴリーごとにまとめたものである．例としては Yahoo!Japan [1] や Open Directory Project [2] などが挙げられる．ディレクトリ型検索エンジンの利点としては，キーワードによらない検索が可能であることや，人手による編集のため索引と要約の信頼性が高いということが挙げられる．その一方で，人手による索引と要約が行われているために，検索対象となる Web サイト群は後述するロボット型検索エンジンと比べてかなり小数であるという欠点がある．

ロボット型検索エンジンとは，自動的に索引付けされた Web ページ群からキーワードを入力することによって Web ページを検索する仕組みである．例としては Google [3] などが挙げられる．ロボット型検索エンジンの利点としては自動的に Web

ページを収集して索引付けしているため，検索対象となるWebページ群はディレクトリ型検索エンジンと比べて多数であるということが挙げられる．その一方で，膨大な量の検索結果が得られてしまうため，求めるWebページを探すことが容易ではないという欠点がある．

そこで本研究では，ロボット型検索エンジンの検索結果など任意のWebサイトを，ディレクトリ型検索エンジンのカテゴリーに自動分類することを目的とする．このような自動分類が可能になれば，ディレクトリ型検索エンジンの欠点である，人手で行っていた索引と要約のコストが削減でき，多数のWebサイトが収集できると考えられる．

ディレクトリ型検索エンジンの情報を用いたWebサイトの分類には，Webサイト中のテキストを利用した方法[4][5]がある．またWebページのタイプへの自動分類のために，Webページ中のリンク数や文字数，タグ数を利用している方法[6]がある．ここでいうタイプとは，掲示板やリンク集など類似した形式で記述された文書のグループであり，カテゴリーとは異なるものである．

本研究では，Webサイト中に現れる画像ファイルやリンク，テキストデータなどの情報から各カテゴリーに特異的に現れる特徴やキーワードの抽出を行う．各カテゴリーごとに特異的に現れる特徴やキーワードが存在すれば，任意のWebサイトをカテゴリーに自動分類することが可能になると考えられる．

2. Webサイト

2.1 Webサイトとは

Webページとは，Webブラウザに一度に表示されるデータのまとまりで，テキストデータやタグ，画像などから構成される．Webサイトとは，関連のあるWebページがリンクによってつながったWebページのまとまりである．本研究では，ディレクトリ型検索エンジンのカテゴリ内のURL群を各トップページとし，それぞれトップページのURLを含み，かつ，リンクでつながっているWebページ群のまとまりを各Webサイトとして定義する．

2.2 Webサイトに含まれる情報

本研究では，Webサイト中のそれぞれのWebページのソースファイルを解析することによって，様々な情報を抽出する．Webサイト S_i に含まれるWebページを $\{P_{i,1},P_{i,2},...,P_{i,N_i}\}$ (N_i:Webサイト S_i のページ数，$P_{i,1}$:Webサイト S_i のトップページ) とすると，Webサイト S_i から抽出する情報は以下の通りである．

- ドメイン

$domain(S_i)$:$\{P_{i,1},P_{i,2},...,P_{i,N_i}\}$ のURLに含まれるドメイン

- URLの深さ

$urldepth(S_i)$: トップページ $P_{i,1}$ のURL中の"http://"を除いた"/"の数

- リンクの深さ

$linkdepth(S_i) = max_n links(P_{i,1}, P_{i,n})$

$links(P_{i,j},P_{i,n})$: $P_{i,j}$ から $P_{i,n}$ への最小リンク数

- 内部間のリンク数

$$inlink(S_i) = \sum_{j=1}^{N_i} inlink(P_{i,j})$$

$inlink(P_{i,j})$: $P_{i,j}$ から $P_{i,k}$(kは任意) へのリンク数

- 外部へのリンク数

$$outlink(S_i) = \sum_{j=1}^{N_i} outlink(P_{i,j})$$

$outlink(P_{i,j})$: $P_{i,j}$ に含まれる $inlink(P_{i,j})$ 以外のリンク数

- 画像数

$$image(S_i) = \sum_{j=1}^{N_i} image(P_{i,j})$$

$image(P_{i,j})$: $P_{i,j}$ 中の".jpg",".gif",".bmp",".png"のカウント数

- 電話番号数

$$tel(S_i) = \sum_{j=1}^{N_i} tel(P_{i,j})$$

$tel(P_{i,j})$: $P_{i,j}$ 中の電話番号数

- メールアドレス数

$$mail(S_i) = \sum_{j=1}^{N_i} mail(P_{i,j})$$

$mail(P_{i,j})$: $P_{i,j}$ 中の"mailto:"のカウント数

- pdfファイル数

$$pdf(S_i) = \sum_{j=1}^{N_i} pdf(P_{i,j})$$

$pdf(P_{i,j})$: $P_{i,j}$ 中の".pdf"のカウント数

- cgiファイル数

$$cgi(S_i) = \sum_{j=1}^{N_i} cgi(P_{i,j})$$

$cgi(P_{i,j})$: $P_{i,j}$ 中の".cgi"のカウント数

- cssファイル数

$$css(S_i) = \sum_{j=1}^{N_i} css(P_{i,j})$$

$css(P_{i,j})$: $P_{i,j}$ 中の".css"のカウント数

- オーディオファイル数

$$audio(S_i) = \sum_{j=1}^{N_i} audio(P_{i,j})$$

$audio(P_{i,j})$: $P_{i,j}$ 中の".mp3",".mpg",".mpeg",".wav",".ram",".rm",".wma",".aif",".asf",".avi",".swf"のカウント数

3. キーワードの抽出

本研究では，2.2節で述べた情報以外に，Webページのソースファイルから単語を取り出すことによってキーワードの抽出を行う．ここでいうキーワードとは，ディレクトリ型エンジンのカテゴリーの特徴を表すような単語，つまり各カテゴリーごとに特異的に現れる単語のことである．単語を取り出す際，日本語の文章は英語の文章のように単語を区切る習慣がないため，漢字仮名混じりの文から単語を抽出しなければならない，という問題がある．そこでソースファイルを形態素解析する必要がある．形態素とは，それ以上小さな単位では意味をなさない単語，または単語の一部のことである．つまり形態素解析とは，

文を形態素に区切って品詞・活用情報を付加する処理のことである．本研究では，形態素解析のシステムとして茶筌[7]を用いて，ソースファイルから日本語の名詞のみを取り出す．

3.1 キーワードの計算方法

本研究では，次の3種類の計算方法でキーワードを計算する．まずカテゴリー中の多くのWebサイトに出現する単語をそのカテゴリーの特徴を表す単語とみなし，カテゴリーiにおける単語wが出現するWebサイトの割合を次の式で計算する．

$$P_i(w) = \frac{\sum_{j=1}^{N_i} E(w, S_{i,j})}{N_i} \quad (1)$$

ここで，N_iはカテゴリーi中のWebサイト数であり，$S_{i,j}$はカテゴリーi中のWebサイトjである．$E(w, S_{i,j})$はWebサイト$S_{i,j}$中に単語wが出現すれば1，それ以外は0をとる．

次に，この値はカテゴリーによって分布が異なり，この値によって分類を行った場合には結果に偏りが生じると推測される．そこで以下の計算を行うことでカテゴリーごとの値の分布を平均0，標準偏差1の正規分布に変換する．

$$Z_i(w) = \frac{P_i(w) - \mu_i}{\sigma_i} \quad (2)$$

ここでμ_iは$P_i(w)$の平均，σ_iは$P_i(w)$の標準偏差である．

最後に，カテゴリーごとに相対的に値の高い単語ほどカテゴリーの特徴をよく表す単語であるとみなし，カテゴリーiにおける単語wのスコア$score_i(w)$を次の式で計算する．

$$score_i(w) = \frac{Z_i(w)}{\sum_{i=1}^{C} Z_i(w)} \quad (3)$$

ここでCはカテゴリー数である．式(3)では相対値をとっているため，もともとZの値が小さいが相対値が極端に高くなる語が存在すると考えられる．このような単語はZの値が小さいのであまり重要ではなくノイズになってしまうので，今回式(3)の計算にあたって各カテゴリーごとにZ値の上位2000語の単語を用いることでこのような単語を除外した．

4. 実　　験

4.1 対象データ

本研究では，Yahoo!Japanに存在する以下のカテゴリーを用いる．

- エンタテイメント>グルメ
- 芸術と人文>美術館・ギャラリー
- 旅行と交通>宿泊施設

カテゴリーの選択については，[8]を参考に観光に関するカテゴリーを対象とした．各カテゴリー中のWebサイト数はグルメ(4440サイト)，美術館・ギャラリー(260サイト)，宿泊施設(7360サイト)である．

4.2 評価尺度

評価尺度を定義するにあたって，まず次に示す値を定義する．

- TP: 正例を正例として判断したWebサイト数
- FP: 正例を負例として判断したWebサイト数
- TN: 負例を正例として判断したWebサイト数
- FN: 負例を負例として判断したWebサイト数

これらの値を用い，評価尺度として適合率，再現率，F1値が次式で定義される．

$$\text{適合率} = \frac{TP}{TP + FP}$$

$$\text{再現率} = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * \text{適合率} * \text{再現率}}{\text{適合率} + \text{再現率}}$$

適合率と再現率はトレードオフの関係にあるが，どちらも考慮する必要がある．そこで，適合率と再現率の調和平均であるF1値を本研究では最も重要な評価基準として用いる．

4.3 設　　定

本研究では，節3.1で示したそれぞれのキーワードの計算方法((1),(2),(3))を用いて未知のWebサイトの分類を行い，それぞれその精度を検証する．

節4.1の各カテゴリーのWebサイトのうちランダムに選んだ90%のWebサイトを用いてキーワードの計算を行い，残りの10%をテストデータとしてキーワードによるカテゴリーへの分類を行う．

分類の際，カテゴリーiとWebサイトjの類似度similarity(i,j)は，例えば(3)の場合，次のように計算される．(1),(2)の場合も同様である．

$$similarity(i, j) = \sum_{k=1}^{M_j} score_i(w_{j,k}) \quad (4)$$

ここでM_jはWebサイトjに出現する単語数，$\{w_{j,1}, ..., w_{j,M_j}\}$はWebサイトjに出現するそれぞれの単語である．

つまりWebサイトに出現する全ての単語に対してカテゴリーごとにキーワードの計算を行い，その和をそのWebサイトの各カテゴリーに対する類似度とする．そのうち最も高い類似度を示したカテゴリーを，そのWebサイトが属するカテゴリーとする．

4.4 結果・考察

キーワードの計算方法	適合率	再現率	F1
P	0.59	0.66	0.62
Z	0.63	0.89	0.73
score	0.91	0.88	0.89

表1　分類精度

カテゴリー	Pの平均	Pの標準偏差
美術館・ギャラリー	0.0244	0.0547
グルメ	0.0068	0.0302
宿泊施設	0.0069	0.0312

表2　カテゴリーごとのPの平均と標準偏差

表1はキーワードの計算方法((1),(2),(3))それぞれによる3カテゴリーへの分類精度を表している．

これより，PよりもZの場合の方がより高い精度を示していることがわかる．

表 2 は各カテゴリーの P の平均と標準偏差である．これから，カテゴリー"美術館・ギャラリー"は比較的平均値が高い，ということがわかる．P では大半の Web サイトがこのカテゴリーに分類されてしまっており，精度が下がってしまったのはこのためであると考えられる．

よって P では分布に偏りがあるため，その分布にしたがって結果も偏ってしまい，精度が下がってしまうということがいえる．しかし，Z では全てのカテゴリーでの値の分布が平均 0，標準偏差 1 の正規分布に変換されているので上記の原因が改善され，精度も向上したということもいえる．

次に，表 1 から Z よりも score の場合の方がより高い精度を示していることがわかる．このことからどのカテゴリーでも高い値を示している単語よりも，各カテゴリーを比較して相対的に値が高い単語の方がカテゴリーに特異的に現れる，特徴を表す単語であるということがいえる．

今回の結果では score の計算方法が最も高い精度を示したが，キーワードによって分類できなかった部分も存在した．このような分類不可能な部分を，Web サイト中のキーワード以外の特徴を用いて分類することでさらなる精度の改善が可能であると考えられる．

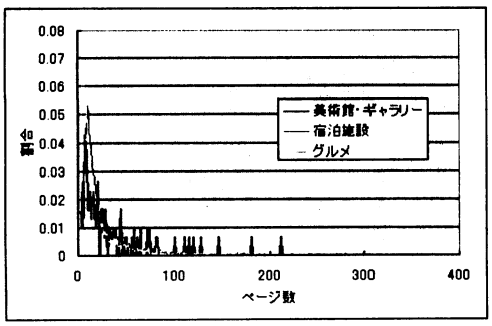

図 1　合計ページ数

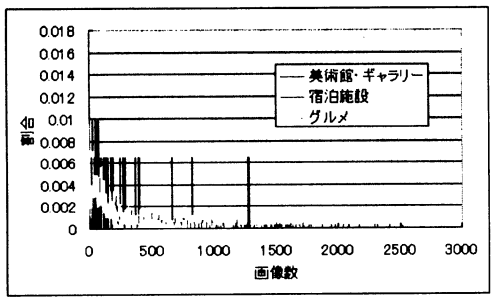

図 2　合計画像数

2.2 節で抽出した特徴のうち，合計ページ数と，合計画像数の分布を図 1,2 に示した．横軸は特徴量，縦軸は全体に対する割合である．図から読み取れるようにカテゴリーによってそれぞれ異なる分布を示しており，これらの特徴を利用することでさらに高い分類精度が得られる可能性があると考えられ，今後検討する予定である．

5. おわりに

本研究では，Web サイトのカテゴリーへの自動分類を行うためにキーワードを用い，その有効性を評価した．その結果，カテゴリーごとに単語が出現する Web サイトの割合を用い，そしてその割合の分布を正規分布に変換することで分類精度を改善できることがわかった．さらに，カテゴリーごとの相対値をとることでより高い分類精度を得ることができることもわかった．そのときの適合率は 91%，再現率は 88%と高い精度を示したが，今回用いたキーワードの計算方法では分類不可能な部分も存在した．そのため現在のキーワードの計算方法の改善が一つの課題として挙げられる．

また Web サイトに現れるキーワードに加えて，キーワード以外の特徴を用いることでキーワードだけでは分類精度が悪かった部分を補うことができるかを今後検討する予定である．

文　献

[1] Yahoo!Japan, http://www.yahoo.co.jp/

[2] Open Directory Project, http://dmoz.org/

[3] S.Brin and L.Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine, Proceedings of the seventh international conference on World Wide Web,1997.

[4] John M. Pierre: Practical Issues for Automated Categorization of Web Sites, ECDL 2000 Workshop on the Semantic Web,2000.

[5] T.Tsukada,M.Washio and H.Matoda: Automatic web-page classification by using machine learning methods, Proceedings of the First AsiaPacific Conference on Web Intelligence,2001.

[6] 久野高志，石田栄美，安形輝，上田修一: Web ページのタイプ判定法，2000 年度日本図書館情報学会春季研究大会発表要綱,2000.

[7] 茶筌, http://chasen.naist.jp/hiki/ChaSen/

[8] 岡本伸之:"観光学入門"，有斐閣アルマ,2001.