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

2007－NL－180 (19)
2007／7／25

Webサイトの階層的なWebディレクトリへの自動分類手法

佐々木 稔† 新納 浩幸†

† 茨城大学工学部 〒316-8511 茨城県日立市中成沢町 4-12-1
E-mail: †{msasaki,shinnou}@mx.ibaraki.ac.jp

あらまし ディレクトリ型の検索サービスはあらかじめ Web ページが項目別にまとめられているので，初心者でも簡単に WWW(World Wide Web) 検索をすることができる．このようなサービスを運営する側は Web ディレクトリへのサイト登録や分類，管理といった作業を人手により行っているため，膨大な Web ページを処理することが困難となる．そのため，我々は人手で行っている Web ディレクトリの管理作業を自動化するシステムの構築を目指している．これまで，サイトの内容語を扱わず，ホームページに記述された meta タグの name 属性値である keyword と description をキーワードとして階層のトップレベルで分類を行い，その結果として分類精度が 82% となり，本文を利用した場合の 55% を大幅に上回る分類性能を得ることができた．本稿では，これまでトップレベルで行っていた分類を拡張し，ディレクトリ階層全体を対象として Web サイトを分類する手法について述べる．階層構造全体を対象とすることで，より現実的で，実用的な Web ディレクトリの構築を行うことが可能となる．階層的な分類においても keyword，description 属性値をキーワードとして利用することの有効性を確かめるために，未分類のデータを利用して実験を行った結果，meta タグのみをキーワードとして利用したシステムは平均 62.7% の分類精度を得ることができた．比較として，meta タグを使わずに HTML 文書の本文を利用した場合の分類結果を求めると 42.3% であった．これより，階層的な分類においても HTML 文書の本文を利用するより meta タグのみを利用した方が有効であることが分かった．また，平均精度が 60% を超えていることから，半自動での Web ディレクトリの構築が可能であると考えられる．

Hierarchical Classification of Web Sites to Web Directory

Minoru SASAKI† and Hiroyuki SHINNOU†

† Department of Computer and Information Sciences, Ibaraki University
4-12-1 Nakanarusawa, Hitachi, Ibaraki, 316-8511, Japan
E-mail: †{msasaki,shinnou}@mx.ibaraki.ac.jp

Abstract A web directory is a directory on the World Wide Web. For example, Yahoo! Directory and Dmoz are well known web directories. Some categories have a lot of web site links across an extensive range of topics. So we browse the categories below to find helpful resources and information. The web directories are created and maintained by human volunteers who are experts in particular categories. So many submissions of registering URLs are delayed due to not selecting the most specific category for them. In our research, we construct a system of automatic classification into a web directory which is maintained by human. In former experiments, the keywords and the description value of the meta tag in HTML documents are very efficient for Web site classification and the effects of the common words cause misclassification of Web sites. In this paper, we describe a classification system for hierarchical web directory structure. Using the whole directory hierarchy, we consider that the system enables to construct a practical and useful web directory. To evaluate the efficiency of this system based on the values of meta tag, we make an experiment on classifying web sites into the Dmoz directory using the web site registered in the Yahoo! directory. As the results of these experiments, the average precision using meta tag is about 62.7% and that using text of HTML document is about 42.3%. The precision using meta tag is higher than using text and we find the efficiency of the meta tag in the hierarchical classification as well as the classification to flat categories.

1. はじめに

ディレクトリ型の検索サービスはあらかじめ Web ページが項目別にまとめられているので，初心者でも簡単に WWW(World Wide Web) 検索をすることができる．このようなサービスを運営する側は Web ディレクトリへのサイト登録や分類，管理といった作業を人手により行っているため，膨大な Web ページを処理することが困難となる．また，充実した Web ディレクトリや個人によるリンク集を構築することも難しい．このため，Web ページの形式や内容からある視点を定め，Web ページを自動分類する研究が盛んに行われている．このような研究には，複数の Web ページにおいて類似した内容をまとめるもの [6]，リンクの参照共起などを分析してつながりの強い文書をまとめるもの [8]，また，Web ディレクトリにおいてリンクの紹介文を利用して分類器を学習するもの [9] が存在する．

この Web ディレクトリに登録された情報は，多くの場合が企業や個人のホームページであり，Web サイト単位で検索をすることができる．しかし，登録された情報が多くなるほど，リンク切れやカテゴリ分けなどの管理が難しくなる．そのため，Web ディレクトリはロボット検索と同様に網羅的な Web の目録を作ることを目的としているが，現状では厳しい登録審査を通った厳選サイトが登録されている．

そこで，我々はこれまで人手で行っている Web ディレクトリの管理作業を自動で行う研究を行っている．オープンディレクトリなどを利用してカテゴリは既に存在するものと考え，まずはディレクトリへのサイト登録を自動的に行うことが課題となる．この課題に対して解決すべき 2 つの問題点を以下に示す．

（1） どのようなページを代表的なサイトとしてディレクトリに登録するか決定すること

（2） 選ばれたサイトを自動的に適切な Web ディレクトリに分類すること

これらの課題に対して，数多くの研究が行われている．初めの問題点については，検索質問からの検索結果を分類対象となる web ページとして分類するものがある [2] [5]．本文の内容やリンク構造を利用して，検索されたページが既存の Web ディレクトリに分類される．2 番目の問題点については，数多くの自動分類方法が提案されている．Web ページをベクトルで表現をして，距離の最も近いカテゴリに分類を行う [1] は，非階層のカテゴリが対象となっている．また，SVM(Support Vector Machine) を利用して階層的なカテゴリに分類を行う実験も行われている [3] [4]．これら実験では，共に分類するデータとして Reuters-22173 が使われているため，ニュース記事の分類が行われている．そのため，既存の Web ディレクトリに対して未登録の Web サイトを階層的に自動分類する実験はあまり行われていない．

我々は，上に示した自動分類の課題に対してこれまでに企業サイトに記載された情報から業種判別を行った [10]．このとき，企業サイトに記載された内容からキーワードを抽出した結果に一般的な単語が数多く含んでいることが原因で，自動分類の精度を大きく下げていることが分かった．そのため，サイトの内容語を扱わず，ホームページに記述された meta タグの name 属性値である keyword と description をキーワードとして分類実験を行った [11]．その結果，トップレベルでの分類精度が 82% となり，本文を利用した場合の 55% を大幅に上回る分類性能を得ることができた．

本稿では，これまではトップレベルで行っていた分類を拡張し，ディレクトリ階層全体を対象として Web サイトを分類する手法について述べる．階層構造全体を対象とするように改良を行うことで，より現実的で，実用的な Web ディレクトリの構築を行うことが可能となる．この階層的な自動分類の性能を評価することにより，現在人手によって行われている分類作業をどの程度まで自動化することができるかについて，評価がしやすくなると考えられる．また，階層的な分類においても keyword，description 属性値をキーワードとして利用することの有効性を示す．

2. Web ディレクトリ

Web ディレクトリは，Web サイトへのリンクをカテゴリ別に分類した階層的なリストである．このような Web ディレクトリの代表的な例として，Yahoo!，Google Directory，goo などの検索エンジンを提供するサイトや Open Directory プロジェクトが世界中のボランティアの協力のもとで作成しているディレクトリ，さらには地域の観光，飲食店情報等を網羅した個人運営によるポータルサイトなどが存在する．ユーザにとってこのような Web ディレクトリが威力を発揮するのは，検索したい分野があらかじめ分かっているときに，素早く欲しい情報を見つけられることにある．

Web ディレクトリに登録されている内容は，そのカテゴリに属する Web サイトのタイトル，URL，概要がひとつの組となって，一覧表示される．図 1 に，Open Directory プロジェクトにより公開されている Dmoz のディレクトリ階層の一部を示す．そこに登録された URL は，多くの場合が企業や個人のホームページであり，1 ページ単位での登録をしているサイトは少ない．ホームページだけであっても，Web ディレクトリの登録や削除などの管理は現在でも人手で行われており，作業の手間がかかってしまう．そのため，できるだけ少ない作業に抑えるために，サイト単位でのディレクトリ設計をしていると考えられる．

現在，Web ディレクトリは上記のような人手による管理の困難さと Web ページの爆発的な増加により作業が追いつかない状態が続いている．また，Google のように，WWW を網羅した精度の高いページ単位での検索が可能となり，現在の検索の主流となっている．そのため，Web ディレクトリはロボット型検索エンジンの検索結果を補完する役割になっている．

3. Web ディレクトリへの自動分類システム

本節では，Web ディレクトリに URL を自動分類するシステムについて述べる．システムは大きく分けて分類モデルの作成部と入力 URL の自動分類部の 2 つからなる．

```
<Topic r:id="Top/Arts/Movies/Titles/1">
  <catid>54803</catid>
</Topic>
<Topic r:id="Top/Arts/Movies/Titles/1/10_Rillington_Place">
  <catid>205108</catid>
  <link r:resource="http://www.britishhorrorfilms.co.uk/rillington.shtml"/>
  <link r:resource="http://www.shoestring.org/mmi_revs/10-rillington-place.html"/>
  <link r:resource="http://us.imdb.com/title/tt0066730/"/>
  <link r:resource="http://online.tvguide.com/movies/database/Movie-Review.asp?MI=22983"/>
</Topic>
<ExternalPage about="http://www.britishhorrorfilms.co.uk/rillington.shtml">
  <d:Title>British Horror Films: 10 Rillington Place</d:Title>
  <d:Description>Review which looks at plot especially the shocking features of it.
  </d:Description>
  <topic>Top/Arts/Movies/Titles/1/10_Rillington_Place</topic>
</ExternalPage>
```

図 1 Dmoz のディレクトリ階層

Fig. 1 Part of the Dmoz directory structure

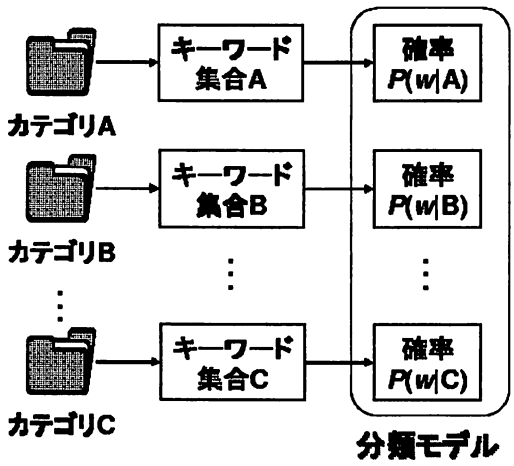

図 2 分類モデルの作成方法

Fig. 2 Construction of Classification Model

3.1 分類モデル作成部

分類モデルの作成部では，同一レベルのカテゴリを対象として，各カテゴリに含まれるキーワードの出現確率を求め，ナイーブベイズ手法に基づく分類モデルを構築する．図 2 に，分類モデルの作成を行う概要を示す．

まず，Web ディレクトリに登録されたデータから，各ディレクトリとそれ以下のサブカテゴリに登録されたサイトからキーワードを抽出する．このとき，抽出するデータとして Web ディレクトリに登録されたリンクのタイトルと紹介文を利用する．Web ディレクトリに登録されたタイトルや紹介文には，Web サイトを重要なキーワードを利用した簡潔な表現が使われている．また，モデル構築を行う際にデータとして利用しやすいことも大きな要因となっている．そのため，登録された Web サイトのデータをダウンロードしなくても有効な分類モデルを構築できると考えられる．

各カテゴリにおいて抽出したタイトルや紹介文のデータ集合に対して，形態素解析を行うシステムである茶筌(注1) を利用して形態素解析を行う．この解析結果で，単語の品詞が名詞 (数，非自立は除く)，片仮名語，未知語，アルファベットをキーワードとして取り出す．これらのキーワードに対して，すべてのディレクトリにおけるキーワードの頻度統計を用いて出現確率を計算する．この統計値とディレクトリ名をラベルとして教師あり学習であるナイーブベイズ手法[7] を利用し，Web ディレクトリの分類モデルとする．

ナイーブベイズ手法はベイズの定理を利用した簡単，かつ強力な分類モデルで，図 2 のカテゴリ A について考えた場合，次の確率を求めることになる．

$$P(A|w) = \frac{P(A)P(w|A)}{P(w)} \tag{1}$$

ここで，$w = (w_1, w_2, \cdots, w_n)$ はカテゴリ A に含まれるキーワードの集合を表す．この式において，条件付き確率 $P(w|A)$ は以下の式で表される．

$$P(w|A) = P(w_1|A)P(w_2|A, w_1)\cdots P(w_n|A, w_1, w_2, \cdots, w_{n-1}) \tag{2}$$

この式を簡単化するために，キーワード $w_1, w_2, \cdots, w_n$ はそれぞれ独立して出現するものと仮定すると，条件付き確率 $P(w_n|A, w_1, w_2, \cdots, w_{n-1})$ は以下のようになる．

$$P(w_j|A, w_1, w_2, \cdots, w_{j-1}) = P(w_j|A) \quad (j = 1, \cdots, n). \tag{3}$$

これにより，式 (1) は以下の簡単な形で書くことができる．

(注1) : http://chasen.naist.jp/hiki/ChaSen/

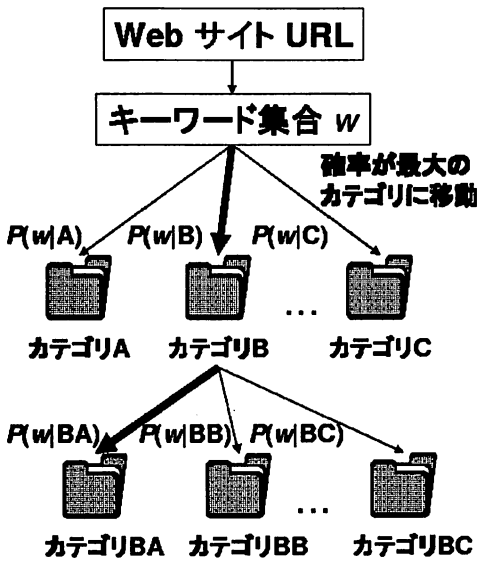

図 3 URL 自動分類の方法
Fig. 3 Classification Method of Web Sites

$$P(A|w) = \frac{P(A)\prod_{i=1}^{n} P(w_i|A)}{P(w)}. \quad (4)$$

この式 (4) を計算することにより，Web サイトをディレクトリに分類するためのモデルを作成することができる．

3.2 URL の自動分類部

URL の自動判定部は，分類したい Web サイトの URL を入力することで，前節で作成した Web ディレクトリの分類モデルを利用して，分類すべきディレクトリの自動的な判定を行う．図 3 に自動判定の概要を示す．

分類したい URL を入力すると，そのアドレスから HTML ファイルをダウンロードする．ダウンロードした HTML ファイルから meta タグの name 属性が keyword(または keywords) と description の値を持つ場合の content 属性値をデータとして抽出する．meta タグは， HTML (Hyper Text Markup Language) 形式で書かれた文書の head タグで囲まれた領域において記述されるもので，HTML 文書で利用される文字コードやキーワード，概要などを埋め込むことができる．これらの属性値を利用するのは，作成したページを端的に表現できるキーワードが選定されていることが多く，また，サイトを簡潔にまとめた概要が記述されているため，有効なデータであることが示されている．これらの属性値から抽出したデータに対して，茶筌を利用して形態素解析を行い，ディレクトリ情報の学習時と同じ品詞属性をもつ単語をキーワード集合として用いる．

得られたキーワード集合に対して分類モデルにより，ディレクトリを判定する．キーワード集合を w とすると，システムはキーワード集合が与えられた時にカテゴリ C に分類される条件付き確率 $P(C|w)$ を計算する．この $P(C|w)$ を最大とするカテゴリ C^* を以下の式で求める．

$$C^* = \operatorname*{argmax}_C P(C|w) = \operatorname*{argmax}_C \frac{P(C)\prod_{i=1}^{n} P(w_i|C)}{P(w)}. \quad (5)$$

このとき，確率 $P(w)$ はカテゴリ C と関係がなく一定値となるので，次のように $P(w)$ を省略して式を簡単にすることができる．

$$C^* = \operatorname*{argmax}_C P(C)\prod_{i=1}^{n} P(w_i|C) \quad (6)$$

このようにして，確率 $P(C|w)$ が最大となるカテゴリが適切なカテゴリであるという分類を行っている．

この分類を階層的に行うために，レベルを再帰的にひとつずつ下げて分類を行うことで，適切なカテゴリの特定を行う．まず，最上位のディレクトリから開始して，どのカテゴリに進むべきかを判定する．このとき，このカテゴリの分類統計はそれ以下の階層に属するすべてのキーワードを含めて分類モデルを作成する．次に，どのカテゴリに進むかを判定する際に，現在いるカテゴリが分類すべき場所である可能性もあるので，下の階層に現在のカテゴリを分類する対象として判定を行う．そこで，現在のカテゴリが選ばれた場合は，そのカテゴリを分類結果とする．この処理を繰り返して，最下層のカテゴリまで移動を行い，それ以上進むことができなくなれば，そのカテゴリを分類結果として返す．

4. Web ディレクトリへの分類実験

本節では，Web ディレクトリに URL を自動的に登録するためのディレクトリ判別実験について述べる．

4.1 使用データ

Web ディレクトリについては，企業の大規模な Web ディレクトリ構造や個人のリンク集など，さまざまな種類の分類データが存在するが，Open Directory Project において作成されたフリーで利用可能な Dmoz [(注2)] を利用する．この Dmoz の Web ディレクトリには数多くの言語で記述されたカテゴリが登録され，日本のカテゴリ ("Top/World/Japanese") 以下を考慮するだけでも 17,519 種類のカテゴリが存在し，それらが階層的につながっている．

この Dmoz をそのまま利用すると問題点が存在する．例えば，ディレクトリの中に "地域" というカテゴリが存在する．その "地域" カテゴリ以下には都道府県名，具体的な内容での分類へと続いている．meta タグの中には具体的な地域名が記載されていることが多く，内容を表すカテゴリに分類されずに，地域以下のカテゴリに分類されやすくなる．また，"オンラインショップ" カテゴリにも問題がある．企業サイトを業種や内容において分類することが目的であるが，具体的な商品を通信販売しているサイトについては，内容により分類されずに "オンラインショップ" に分類される傾向がある．これら特定のカテゴリへの誤分類の問題に関しては，文書内容とは異なるカテゴリとなるため，改めてこれらのカテゴリへの分類を行い，重

(注2)：http://dmoz.org/

表 1 分類対象となるトップカテゴリ一覧
Table 1 A set of the first level categories in Dmoz

トップカテゴリ	データ数
アート	15552
キッズとティーンズ	1345
ゲーム	2260
コンピュータ	3657
スポーツ	9214
ニュース	1501
ビジネス	21630
レクリエーション	8818
健康	5014
各種資料	2788
家庭	2237
社会	8970
科学	2190

表 2 実験で利用したカテゴリ一覧
Table 2 A set of categories used in our experiment

カテゴリ
エンターテインメント/映画，ビデオ
エンターテインメント/音楽
エンターテインメント/コミックとアニメーション
コンピュータとインターネット/インターネット
コンピュータとインターネット/ソフトウェア
コンピュータとインターネット/ハードウェア・プログラミング
趣味とスポーツ/スポーツ

複して分類を行うことで対応が可能であると考えられる．その対応は今後の課題とする．

以上のような理由により，本稿で行う実験は，日本のカテゴリから “地域” と “オンラインショップ” を除いたトップカテゴリ以下の階層を対象として分類を行う．分類の対象となるトップカテゴリの一覧を表 1 に示す．

4.2 分類実験

Dmoz に登録された Web サイトのデータから学習した分類モデルを利用して，そこに含まれていない新しい URL から分類すべきディレクトリの自動判別を行った．テストデータには，表 2 に示す Yahoo! カテゴリ(注3) に登録された 7 種類のカテゴリに含まれる，Dmoz に登録されていない各 100 件の URL を利用した．

分類する URL から HTML 文書をダウンロードし，その中に必要とする meta タグの keyword と description の name 属性値が含まれていれば，その内容を抽出する．抽出したデータに対して，分類モデルの作成時と同様に形態素解析を行い，片仮名文字列，未知語，アルファベットや名詞 (数，非自立は除く) をキーワードとして利用する．このとき，キーワードの頻度に重みを加えるために，出現頻度を 2 乗した値をキーワードの重要度とする．このキーワード統計に対してナイーブベイズ手法を利用して，移動するカテゴリの判定を行う．この判定処理をトップカテゴリから再帰的に階層を移動し，カテゴリの移動ができなくなったとき，そのカテゴリを分類結果として決定する．

この分類システムの有効性を評価するため，比較実験として Web サイトのトップページからダウンロードした HTML 文書に対して，meta タグを使わず，本文のみからキーワードを抽出した場合についての実験を行った．この比較実験の結果と meta タグを利用した場合の分類性能を比較し，評価を行う．比較実験については，本文のキーワードの重み付けとして，Web ページに出現するキーワードの頻度をそのまま利用することとした．

4.3 実験結果・考察

本実験を行った結果を表 3 に示す．表 3 は，テストデータが元のカテゴリ名とほぼ同じ Dmoz のカテゴリに分類された文書数を，meta タグのみを利用した場合と本文のみを利用した場合についてそれぞれ表している．この表より，meta タグのみを利用したシステムは平均 62.7% の分類精度を得ることができた．「映画，ビデオ」，「音楽」，「スポーツ」といったカテゴリは他のジャンルとの関係が少ないこともあり，平均以上の精度が得られた．全体的に 60% を超える分類性能があれば，自動で分類した結果を確認し，誤っていれば修正を行う半自動的な Web サイトの管理も可能ではないかと考えられる．ただ，「ソフトウェア」に関しては様々な種類のソフトウェアが存在するため他のカテゴリとの関わりが強く，分類精度が低くなっている．また，種類が多さはカテゴリが存在しない場合もあり，以下の表 4 に示す実験結果の例では，Dmoz のカテゴリにオープンソースに関するカテゴリが存在していないことで，誤った判定結果となっている．

表 4 分類できなかったサイトの例
Table 4 Example of the web site misclassified

URL	http://www.opensource.jp/
Yahoo のカテゴリ	トップ/コンピュータとインターネット/ソフトウェア/オープンソース/団体
Dmoz のカテゴリ	World/Japanese/コンピュータ/ プログラミング/国際化

次に，meta タグを使わずに本文のみからキーワードを抽出した場合と比較すると，階層的な分類においても meta タグを利用する方が分類性能が良かった．Web ページからキーワードを抽出する際，分類に有効なキーワードだけではなく，ある程度の幅を持つ分野で出現するキーワードが数多く抽出されていたためだと考えられる．分類モデルの作成時におけるキーワード抽出方法も含めて，紹介文や meta タグ以外で有効なキーワードの抽出を行う必要がある．また，効果的な分類を行うためには分類にはあまり関係のない一般的なキーワード，すなわちアウトライヤーの除去を行うことが有効ではないかと考えられる．カテゴリ間の関係やキーワード間の関係などを利用して，分類が効果的に行われるキーワードのみを残すことが今後の課

(注3) : http://dir.yahoo.co.jp

表 3 meta タグと本文を利用したときの実験結果
Table 3 Experimental Results using meta tag and text

トップカテゴリ	meta タグ	本文のみ
映画，ビデオ	70	36
音楽	68	60
コミックとアニメーション	60	23
インターネット	62	62
ソフトウェア	49	46
ハードウェア・プログラミング	60	32
スポーツ	70	37
合計	439	296
分類精度 (%)	62.7	42.3

題である．

5. おわりに

本稿では，入力された Web サイトの URL に対して，ディレクトリ階層全体を分類対象として自動的に分類するシステムの提案を行った．階層的な分類においても keyword，description 属性値をキーワードとして利用することの有効性を確かめるために，未分類のデータを利用して実験を行った．その結果，meta タグのみをキーワードとして利用したシステムは平均 62.7% の分類精度を得ることができた．比較として，meta タグを使わずに HTML 文書の本文を利用した場合の分類結果を求めると 42.3% であった．これより，階層的な分類においても HTML 文書の本文を利用するより meta タグのみを利用した方が有効であることが分かった．また，全体的に 60% を超える分類性能が得られたことから，自動で分類した結果を確認し，誤っていれば修正を行う半自動的な Web サイトの管理が可能であると考えられる．

今後は，より分類精度を高めるために，meta タグに記述された単語の分析を行い，本文中のキーワードとの関連性を導き出し，より有効なキーワード抽出を実現することが課題である．また，分類にはあまり関係のない一般的なキーワードを除去し，分類を効果的に行うことができるキーワードを残すフィルタリングを行って，全自動で分類を行うことができるシステムを目指したい．

文　献

[1] C. Chekuri, M. Goldwasser, Prabhakar Raghavan, and E. Upfal. Web search using automatic classification. In *Proceedings of WWW-96, 6th International Conference on the World Wide Web*, San Jose, US, 1996.

[2] Hao Chen and Susan Dumais. Bringing order to the web: automatically categorizing search results. In *CHI '00: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 145–152, New York, NY, USA, 2000. ACM Press.

[3] Susan T. Dumais and Hao Chen. Hierarchical classification of Web content. In Nicholas J. Belkin, Peter Ingwersen, and Mun-Kew Leong, editors, *Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval*, pages 256–263, Athens, GR, 2000. ACM Press, New York, US.

[4] Pei-Yi Hao, Jung-Hsien Chiang, and Yi-Kun Tu. Hierarchically svm classification based on support vector clustering method and its application to document categorization. *Expert Syst. Appl.*, 33(3):627–635, 2007.

[5] Krishna Kummamuru, Rohit Lotlikar, Shourya Roy, Karan Singal, and Raghu Krishnapuram. A hierarchical monothetic document clustering algorithm for summarization and browsing search results. In *Proceedings of the Twelth International World Wide Web Conference*, New York, May 2004.

[6] 石田 栄美, 久野 高志, 安形 輝, 野末 道子, 上田 修一. 内容的なまとまりをもつＷｅｂページ群の自動判定. 1999 年度三田図書館・情報学会研究大会発表論文集, 三田図書館・情報学会, 1999.

[7] 北 研二. 確率的言語モデル. 東京大学出版会, 1999.

[8] 原田 昌紀, 風間 一洋, 佐藤 進也. 参照共起分析の web ディレクトリへの適用. 情報学基礎研究会, 情報処理学会, 2001.

[9] 谷津 哲平, 新納 浩幸, 佐々木 稔. Web ディレクトリを用いた検索ナビゲーション. 言語処理学会第 *11* 回年次大会論文集, pages 1022–1025, 2005.

[10] 佐々木 稔, 新納 浩幸. 文書分類手法を用いた企業 web サイトからの業種分類. 言語処理学会第 *12* 回年次大会論文集, pages 352–355, 2006.

[11] 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, pages 895–898, 2007.