Webディレクトリを用いた検索ナビゲーション

谷津哲平
茨城大学大学院
理工学研究科

新納浩幸
茨城大学工学部
システム工学科

佐々木稔
茨城大学工学部
情報工学科

1 はじめに

Web ディレクトリは，人知を用いて分類されたサイトの索引集である．似ているテーマを扱うサイトが紹介文とともに，ひとつのカテゴリを形成し，木構造をなしている．これは Web 全体を整理したものと捉えることで，知識の大系を示しているといえる．そこで近年，その Web ディレクトリを知識の大系とみたときの利用方法がいくつか提案されている [1][2]．本研究では Web ディレクトリの利用方法として．検索ナビゲーションへの利用を検討する．

検索ナビゲーションとは，検索エンジンにおいて，利用者の詳しくない分野などの検索を補助する機能である．一般的な補助の機能には，関連するサイトを示す，利用者のさがしている分野を絞りこむなどがあり，最近では目的の情報にたどり着くまでの過程が注目されている．

本研究では「より周辺情報の得られる検索」を考え，検索キーワードが Web ディレクトリのどの位置に属するかを示すナビゲーションを考える．そこで，Web ディレクトリ上のあるカテゴリ名と，そこに属するサイトの紹介文に出現する単語には，何らかの関係が見られることに注目した．サイトの紹介文から単語を抽出し，そのカテゴリ名をその単語のクラスとし，それを訓練データとして分類器におけるクラスの定義に用いる．この機械学習により，Web ディレクトリという大系における位置のナビゲーションを実現する．

2 Webディレクトリ

代表的なディレクトリ型検索エンジンの中にボランティアで運営されている Open Directory Project(ODP) の検索サイト dmoz[1]がある．ODP の Web ディレクトリのデータは無償で提供されており，これらは Google ディレクトリを含む多くの検索エンジンで利用されている．そこで本研究では，ODP の Web ディレクトリのデータを利用することにし，CustomDir[2] より入手した．CustomDir は ODP のデータを用いたディレクトリサイト開発支援ツールやデータの総称であり，株式会社スプラインによって無償で提供されている．

以下に本研究で使用した Web ディレクトリの構成の一部を示す．これは『Top/World/Japanese/コンピュータ/ ソフトウェア/』以下の 2 階層分である．実際には最も深かいところで，この下位に 3 階層分あり，利用したカテゴリの総数は 138 個である．また，カテゴリ名の後にある括弧内の記号は，我々の与えたクラス名である．

- ソフトウェア (0)
  - ダウンロード (A)
  - 翻訳 (B)
  - インターネット (C)
    - ネットワーク管理 (CA)
    - クライアント (CB)
    - ホームページ作成・管理 (CC)
    - サーバー (CD)
  - グラフィックス (D)
    - 3D(DA)
    - イメージ編集 (DB)
    - Web_デザイン (DC)
    - ベクタベース (DD)
    - DTP(DE)
  - デスクトップカスタマイズ (E)
    - 壁紙 (EA)
    - アイコン (EB)
    - スクリーンセーバー (EC)
  - データベース (F)
    - Access(FA)
    - Oracle(FB)
    - ファイルメーカー Pro(FC)
    - PostgreSQL(FD)
    - 4th_Dimension(FE)
    - DB2(FF)
  - オペレーティングシステム (G)
    - Windows(GA)
    - DOS(GB)
    - TRON(GC)
    - Unix(GD)

[1]http://www.dmoz.org/
[2]http://www.customdir.net/

- Plan9(GE)
- BeOS(GF)

- 教育 (H)
- プレゼンテーション (I)
  - PowerPoint(HA)
- ワードプロセッサ (J)
  - Word(JA)
  - 一太郎 (JB)
- エディタ (K)
  - Emacs(KA)
- 表計算 (L)
  - Excel(LA)
- 財務・会計 (M)
- オフィススイート (N)
  - OpenOffice.org(NA)
  - Microsoft_Office(NB)
- グループウェア (O)
  - Wiki(OA)
- ファイル管理 (P)
- ビジネスグラフィックス (Q)
- 産業別 (R)
  - 建設 (RA)
- セキュリティ(S)
  - ウィルス対策 (SA)
  - 暗号 (SB)
  - ファイアウォール (SC)

3 検索ナビゲーション

3.1 提案するナビゲーション

本研究で提案する検索ナビゲーションは以下の手順に従う.

1. 検索キーワードを入力する
   例: 「コンピュータウィルス」を入力

2. 第一階層から予想されるカテゴリ名を返す
   例: 第 1 候補 セキュリティ
   第 2 候補 インターネット
   第 3 候補 データベース

3. さらに絞りこむ場合は，第二階層からカテゴリ名を返す
   例: 第 1 候補 ウィルス対策
   第 2 候補 ファイアウォール
   第 3 候補 暗号

このときディレクトリのツリーを図で表示して，入力した単語群が，全体の大系の中のどの位置にあるかを示し，その周辺にはどのようなカテゴリが存在するかを目で確認できるようにする．この検索の過程により周辺情報やもっと興味のあるものを見つけられるようにする.

最近のディレクトリ型検索エンジンには，検索窓がついているものも存在する．例えば Google ディレクトリでは，ページ検索された中から登録されているサイトを返し，dmoz では，カテゴリ名のみを検索対象としている．これらはディレクトリ型側からの検索補助であるが，本研究の趣旨はロボット型検索エンジンにおける Web ディレクトリの利用である．登録してあるサイトを返すのではなく，ナビゲーションの観点から入力した単語群がどのような分野に属するかを返すものである.

3.2 訓練データの作成

前項のナビゲーションシステムを，クラス分類を用いることで実現する．そのための訓練データは以下のように作成した.

まず，Web ディレクトリの各カテゴリ内のサイト紹介文から名詞，未定義語を抽出し，抽出した単語にカテゴリ名に対応したクラスと出現回数を付与する．次にこれらを検索の階層に応じて，ひとつのファイルに統合することで訓練データを作成した (図 1).

訓練データは全部で 20 個作成した．第一階層用の訓練データには「ソフトウェア」以下の全ての単語が含まれた 1 つが必要である．また，第二階層用のものには，第一階層のカテゴリ (クラス A〜S) 毎にそれぞれその下位階層の単語がすべて含まれている 19 個の訓練データを用意した.

そして，この訓練データを使って，入力された検索キーワードのカテゴリ名の類推を行うが，このカテゴリ名 (クラス) を決定する手法については次項に示す.

なお，サイトの紹介文や，入力キーワードから名詞・未定義語を抽出するための形態素解析には，京都大学の JUMAN を用いた.

3.3 Naive Bayes の利用

クラス分類問題とは，分類対象とそのクラスの組からなる学習事例から，未知の分類対象をクラスに分類する問題である．例えば，分類対象 $x = \{f_1, f_2, \cdots, f_n\}$ の分類先のクラスの集合を $C = \{c_1, c_2, \cdots, c_m\}$ とすると，この分類問題は $P(c|x)$ の分布を推定することで解決でき, x のクラスは以下の式で求まる.

$$c_x = arg \max_{c \in C} P(c|x)$$

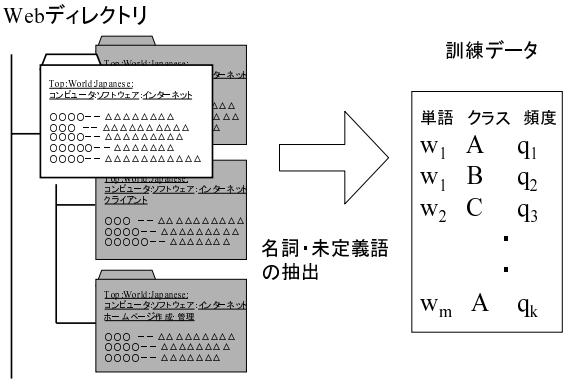

図 1: 訓練データの作成

これは各クラスにおいて x がそのクラス c に属する確率を求め，最も確率が高くなるときの c が，x のクラスということである．この式は Naive Bayes 法を用いることで次の式になる [3].

$$c_x = arg\max_{c \in C} P(c) \prod_{i=1}^{n} P(f_i|c)$$

本研究において，分類対象は Web ディレクトリの紹介文に出現する単語であり，クラスはその紹介文の書かれているカテゴリ名に対応する．そして未知の分類対象が，検索エンジンで利用者が入力する検索キーワードに対応する．

分類問題への変換は以上であり，前項に示した訓練データと Naive Bayes 法を用いて検索キーワードのクラスの推定を行う (図 2).

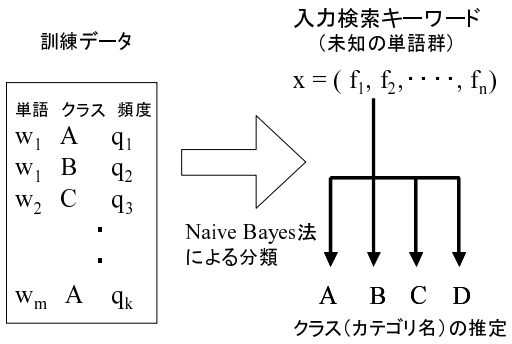

図 2: クラスの推定

4 実験

4.1 評価方法

前項の方法によって作成したシステムを評価するため，次のような交差検定を行った．

図 3 のようにひとつのカテゴリ内には複数のサイトが紹介されている．この中のひとつの紹介文中の単語を訓練データから削除し，その抜いた単語群を入力キーワードとしたときの結果を調べた．紹介文の属するクラスが返れば正解である．これをカテゴリ内のすべての紹介文ごとに行う．

対象のカテゴリは，無作為に次の 6 つを選択した．第一階層より「翻訳」「表計算」「財務・会計」を，第二階層より「ホームページ作成・管理」「Web_デザイン」「ウィルス対策」を選択した．この第二階層のカテゴリに関しては，第一階層 (クラス A〜S) の類推と，それぞれの第二階層カテゴリ内での類推 (例えば「ウィルス対策」なら (SA)〜(SC) の類推) を行った．

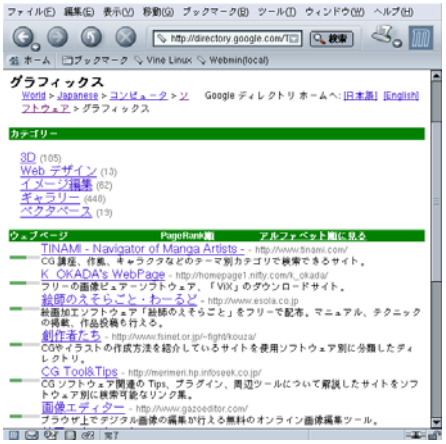

図 3: カテゴリ内の紹介文

4.2 結果

第一階層の「翻訳」のカテゴリでは，全 7 つの紹介文において 5 つが $P(c|x)$ の最も高いものに正解のクラスを返した．同様に「表計算」では，4 つ中 1 つ,「財務・会計」では 18 個中 9 つが 1 番目に正解のクラスを返した．また,「ホームページ作成・管理」では 31 個中 30 個,「Web_デザイン」では 15 個中 11 個,「ウィルス対策」では 12 個中 10 個が正解のクラスを返した．

一方，第二階層単独では「ホームページ作成・管理」が 31 個の紹介文の中 26 個,「Web_デザイン」では 15 個中 13 個,「ウィルス対策」では 12 個中 12 個全てが正解のクラスを返した．

クラス推定において確率 $P(c|x)$ の最も高いクラスに，正解のクラスが選ばれなかったものでも，2 番目や 3 番目までに選ばれていたものが多かった．表 1，表 2 は，何パーセントの割合で正解のクラスが選ばれたかを 2 番目以内，3 番目以内に選ばれた確率と共に示したものである．この 2 番目や 3 番目に選ばれたクラスはナビゲーション過程のカテゴリ候補の提供に利用できる．

表 1: 第一階層の正解率 (%)

カテゴリ名	1 位	2 位以内	3 位以内
翻訳	71.4	85.7	85.7
表計算	25.0	75.0	75.0
財務・会計	50.0	77.7	83.3
ホームページ作成・管理	96.8	96.8	100.0
Web_デザイン	73.3	86.7	93.3
ウィルス対策	83.3	91.6	100.0

表 2: 第二階層の正解率 (%)

カテゴリ名	1 位	2 位以内	3 位以内
ホームページ作成・管理	83.9	96.8	100.0
Web_デザイン	86.7	100.0	100.0
ウィルス対策	100.0	100.0	100.0

※第一階層から通しての正解率「ホームページ作成・管理」80.6%,「Web_デザイン」60.0%,「ウィルス対策」83.3%

5 考察

表 1，表 2 からみて，正解のクラスが上位に選ばれていることが分かる．また，第二階層の方が第一階層よりも，さらに良い結果が得られていることが分かる．その理由としては，第二階層の紹介文はより単語が具体的になり，絞りこまれている状態であるためと考えられる．

本実験結果から，Web ディレクトリは検索ナビゲーションに利用できると考えられる．検索ナビゲーションを実際に行うには，Web ディレクトリに相当する知識体系を予め作っておく必要があるが，既存の知識体系である Web ディレクトリを流用できることは大きな利点である．また Web ディレクトリのディレクトリの位置を特定するための規則の構築も通常困難であるが，本手法では紹介文を訓練データにして機械学習により，その規則の構築を行っているので，規則の構築のコスト削減の面でも大きな利点を持つ．

また，この実験によって，誤って一番目に選ばれたクラスを調べると正解のクラスに似ている分野のカテゴリであることも分かった．これを利用して，新しいカテゴリを追加する検討も行える．

6 おわりに

本論文では Web ディレクトリをシステム側の知識の大系として，検索ナビゲーションに利用した．

最近の検索エンジンは，実際，返すページの多さよりも，満足するサイトにどれだけ早くたどり着けるかや，周辺情報が拾いやすいかなどが重要になっている．その中で，本検索ナビゲーションから得られる大系の中での現在位置を知ることは，情報の検索過程において新しい発見ができるものであると考える．

今後の課題としては，対応するカテゴリの範囲を広げると共に，Web ディレクトリの他の利用方法を考えていきたい．

参考文献

[1] Celina Santamaria,Julio Gonzalo,and Felisa Verdejo： “Automatic Association of Web Directories with Word Senses”, Conputational Linguistics,Volume 29,Number 3,pp.485-502 (2003).

[2] 木村 文則, 前田 亮, 吉川 正俊, 植村 俊亮： “Web ディレクトリの階層構造を利用した言語横断情報検索”, 第 14 回データ工学ワークショップ (2003).

[3] Tom M.Mitchell,Thomas Michell ： “Machine Learning(Mcgraw-Hill Series in Computer Science)”, (1997).