階層型知識体系を用いたWWW情報の自動カテゴリ推定方法[1]

5Z－6

村本 達也 鷲崎 誠司[2]
NTT 情報通信研究所[3]
E-mail : {muramoto, suzaki}@isl.ntt.co.jp

1 はじめに

近年のインターネットの爆発的普及に伴い，WWW(World Wide Web) 上には無数の情報が存在する．この大量の情報の中から利用者に有用な情報の提供を可能にする情報フィルタリング手法が注目されている．情報フィルタリング手法は利用者の興味，趣味，嗜好等を基に情報を提示する方法が多いが，従来方法では利用者に陽に入力を要求するか，閲覧した情報内の特徴語句を抽出することにより推定していた．これに対して我々は，閲覧した情報を既存の階層型知識体系に対応させることにより，利用者興味の推定の精度の向上を目指している．本稿では，これを実現するためのWWW情報の自動カテゴリ推定方法とその評価について述べる．

2 階層型知識体系

WWW 上における階層型知識体系とは，Yahoo Japan[1] やNTT DIRECTORY[2] に代表される登録型のディレクトリ型検索サービスのように，構造が多段階の階層をなしており，各階層が上位概念・下位概念の関係を保持し，階層の各カテゴリにはそれに対応する情報 (WWW 情報) が保持されているものとする．以下，階層型知識体系の節点をカテゴリ，あるカテゴリから根のカテゴリまでのカテゴリの列をパス，カテゴリに保持されているWWW 情報を情報と呼ぶ．また，情報の分類項目は階層型知識体系のパスで表現する (例 スポーツ：サッカー：W杯)．

3 自動カテゴリ推定方法

3.1 機能構成

本方法は，(1)WWW 情報のテキストから形態素解析を用いた単語の抽出，(2) 抽出した単語の階層型知識体系への対応付けによるカテゴリの推定，により実現されている．機能構成の外観を図 1 に示す．

処理 (1) では，当該情報からだけではなく，当該情報が参照している情報（リンク先の情報）からも同様に単語を抽出する．これは，当該情報がフレーム組の情報であったり，イメージ情報のみであったりして，テキスト情報を全く得られない，または得られても特徴的な語句を抽出できないという事が考えられるからである．より多くの参照情報から単語を抽出することで，より有用な単語の取得が期待できるが，限界効用と計算効率を考えて，5 個の参照情報から抽出することにする．抽出した単語の中から特徴語句になりやすい普通名詞，固有名詞，形容動詞を採用し，出現度数の多いものを使用する．

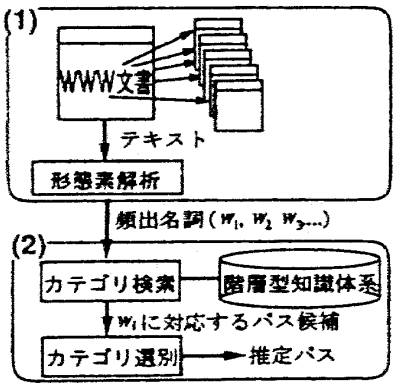

図 1 ：機能構成の外観

処理 (2) におけるカテゴリ推定では，図 1 に示すカテゴリ検索装置を用いる．これは，従来のディレクトリ型検索サービスが持つ機能の，検索語句を投入すると，その語句に対応した情報とその情報が格納されている階層型知識体系のパスの組をリストアップする機能である．これを利用して，先に抽出した単語を階層型知識体系に対応させることで，分類先を推定する．具体的な方法は，始めに複数の単語をカテゴリ検索装置に投入して，投入語句に対応するパスと頻度を得る．次に検索で得たパスの頻度の和を各パスについて計算する．そしてパスを頻度の和でソートして上位を採用することにより，推定する．即ち，ここで推定されたパスは，当該情報とその参照情報中の頻出単語を含む情報を多く保持していることになる．このような方法で処理することにより，もし処理 (1) で抽出された語句に，特徴的でないもの（例：更新）が抽出されたしても，この語句に対応するパスは多種になりやすく，頻度の和でソートすると下位になる可能性が高くなる．また，的外れな語句が抽出されたしても，この語句に対応するパスと他の語句に対応するパスは重複しにくいので，頻度の和でソートすると下位になる可能性が高くなる．

---

[1] An Automatic Classification of the WWW Information Using a Hierarchal Knowledge-Base

[2] Tatsuya MURAMOTO, Seiji SUSAKI

[3] NTT Information and Communication Systems Laboratories

3.2 実行例

幾つかのWWW情報が含まれるパスを推定した結果を図2に示す．これは階層型知識体系にNTT DIRECTORYを用いた場合であるが，各パスは必ずしも互いに素ではないので，図中「JavaFAQ，ＩＣカード」の様に複数のパスに分類される情報も存在する．

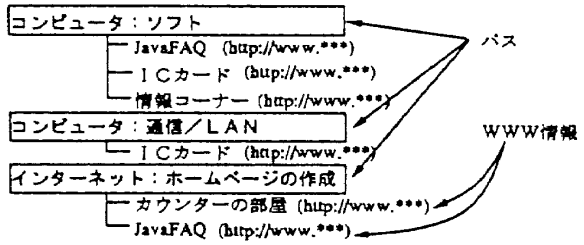

図2：実行例

4 評価

実験には1997年9月にNTT DIRECTORYに登録された情報の中から無作為に抽出した51件を用いた．階層型知識体系として，NTT DIRECTORYとYahoo Japanを用いた．評価尺度として，

$$\text{適合率} = \frac{\text{正解パス数}}{\text{採用したパス数}}$$

$$\text{正当率} = \frac{\text{正解を得た情報数}}{\text{全情報数(51件)}}$$

を用いる．

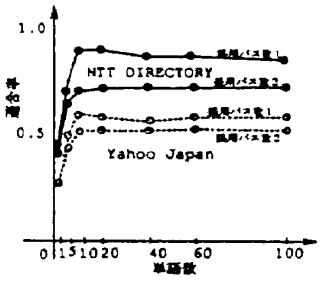

図3：採用単語数の妥当性

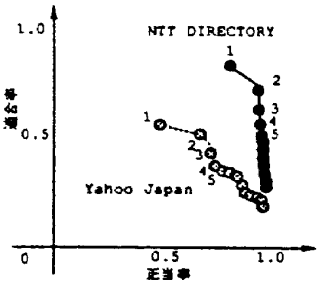

図4：採用パス数の妥当性

図3は形態素解析で抽出された頻出単語の上位k個（横軸）を用いてパスを検索し，上位1個又は2個のパスを採用した時の適合率（縦軸）である．この図から分かるように，NTT DIRECTORYとYahoo Japanのどちらを階層型知識体系として利用したときも，また採用するパスが1個でも2個でも，採用する単語の個数が10個を越えると適合率の上昇は見られない．よって適合率から判断すると上位10個の採用が妥当である．

図4は形態素解析で抽出された頻出単語の上位10個を検索語句として投入し，検索結果の上位k個のパスを採用したときの正当率（横軸）と適合率（縦軸）のプロットである．またグラフ中の数字はkにあたる．この図から分かるように，NTT DIRECTORYとYahoo Japanのどちらを階層型知識体系として利用したときも，採用パス数を増やしていくと，(1.0, 0.0)に収束していく事が分かる．また，正当率と適合率に同じ重みを持たせて目的関数を 正当率＋適合率 とすると，これを最大にするのは$k = 2$の時である．よって，検索結果からパスを2個採用するのが最も有用であるといえる．

上記の実験で得たパラメタである，抽出した頻出名詞の上位10個を検索語句として投入し，検索結果の上位2個のパスを採用する事により，NTT DIRECTORYで92%，Yahoo Japanで73%の正当率を得た．Yahoo Japanはカテゴリの分類が細かく，総カテゴリ数がNTT DIRECTORYの5倍以上保持しているため，本方法では良い正当率が得られなかったものと推測される．しかし，Yahoo Japanを階層型知識体系として用いた場合，NTT DIRECTORYを用いた場合よりも詳細に分類されるため，より詳細な情報を得ることが出来る．また，これが両者の構造の善し悪しの基準となるものではない．

5 まとめ

本稿では，既存の階層型知識体系を用いたWWW情報の自動カテゴリ推定方法を提案した．また，階層型知識体系としてNTT DIRECTORY, Yahoo Japanを用いて実験を行い，高い正当率でカテゴリが推定できることを確認した．

本方法は，情報フィルタリング手法の利用者興味の自動抽出を念頭に置いているが，現在手作業で行われているディレクトリ型検索サービスの情報の登録の支援や，ロボット型検索サービスの検索結果の再構築等への応用が考えられる．

今後の課題として，単語を抽出する時に，出現度数ではなく$tf \times idf$[3]で正規化した頻度を用いた場合との比較等を考えている．また，再現率を計測し，適合率とのバランスをも考察する予定である．

謝辞

本システム実現の為にYahoo Japanのデータの利用を承諾していただいたソフトバンク(株)の方々に感謝いたします．

参考文献

[1] http://www.yahoo.co.jp/

[2] http://navi.ntt.co.jp/

[3] G. Salton : Automatic Text Processing, *Addison-Wesley*, (1989).