E-050

Webを用いた概念の形成とカテゴリ誘導
Formation of a Concept and Category Guidance using Web

堀田 真次 † 中村 正昭 † 渡部 広一 † 河岡 司 †
Shinji Hotta Masaaki Nakamura Hirokazu Watabe Tsukasa Kawaoka

1. はじめに

人にやさしく，使いやすい検索エンジンの実現を目指すために，ユーザの入力語の意味を理解し，入力語の意味に適合するカテゴリに誘導する必要があると考える．現在のディレクトリ型検索エンジンでは，語と語の表記一致によりカテゴリ誘導を行っている．しかし，入力語とカテゴリが表記は違っても，意味が類似している場合には，そのカテゴリへ誘導することができない．本稿では，ユーザの入力語とカテゴリの意味をそれぞれ Web から機械的に取得し，意味を比較することにより適合するカテゴリに誘導するための手法を提案する．本稿におけるカテゴリとは，Web ページを分類する際に基準となる分類項目を指す．

2. 語の意味の比較によるカテゴリ誘導

表記一致による誘導では，同義語辞書などを用いることにより，例えば「買い物」と「買物」といった表記の違いの問題は解決することができるが，以下のような誘導を行うことはできない．

・ZZ-R ⇒ オートバイ，メーカーとモデル

・誕生石 ⇒ ジュエリー，アクセサリー

ユーザの入力語が「オートバイ」あった場合には,「オートバイ」というカテゴリに誘導することはできるが，上記の例のように,「ZZ-R」というオートバイの機種名から「オートバイ」というカテゴリに誘導することはできない．このように,「ZZ-R」からカテゴリ「オートバイ」,「メーカーとモデル」に誘導を行うためには，語の表記からではなく語の意味による比較を行う必要がある．

3. 概念の定義

語と語の意味の比較を行うために，語を概念とし，概念を属性と重みの対の集合で表現する．属性とは概念の意味特徴を表す単語であり，概念と属性の関係の深さを定量化するための指標が重みである．つまり，重みが大きい属性ほど概念との関係が深くなる．ある概念 A を以下のように定義する．

$$\text{概念 } A = \{(a_1, w_1), (a_2, w_2), \cdots, (a_N, w_N)\} \quad (1)$$

式 1 を用いることにより，概念と概念の関係の深さを計算することができ，意味的な比較が可能となる．

† 同志社大学大学院　工学研究科
Graduate School of Engineering,Doshisha Univ.

4. TF・IDF による重み付け

属性の重み付け手法として，情報検索の分野で一般的に用いられるＴＦ・ＩＤＦ [2] を用いる．属性 a_j の文書 D_i における重みを $w_i(a_j)$ とすると，$w_i(a_j)$ は以下の式で定義することができる．

$$w_i(a_j) = tf_i(a_j) \times idf(a_j) \quad (2)$$

$$tf_i(a_j) = f_{ij} \quad (3)$$

$$idf(a_j) = 1 + \log(N/n_j) \quad (4)$$

f_{ij} は a_j の文書 D_i における出現頻度，N は総文書数，n_j は a_j が現れる文書数である．

ＴＦ・ＩＤＦによる重み付けとは，単語に頻度と網羅性に基づいた重み付け手法である．$tf_i(a_j)$ は頻度情報を用いて適切な情報を集めたものであり，$idf(a_j)$ は特定性情報を用いて特徴付けた情報を集めたものである．

5. カテゴリ誘導

本稿では，カテゴリ誘導を行うために，カテゴリとユーザの入力語（検索質問文）をそれぞれ概念とし，カテゴリと検索質問文の関係の深さを，ベクトル空間モデル [1] を用いた類似度計算により求める．そして，類似度の高いカテゴリへとユーザを誘導する．

5.1 カテゴリの構造

ディレクトリ型検索エンジンでは，Web ページの情報を内容や目的などで分類している．Web ページの分類項目をカテゴリと呼ぶ．本稿では，カテゴリを Yahoo! JAPAN [3] のカテゴリ構造を参考に生成した．生成したカテゴリ構造の一部を表 1 に示す．

表 1: カテゴリ構造

カテゴリ名	子カテゴリ	親カテゴリ	階層
教育	大学	-	1
大学	学生生活	教育	2
自動車	オートバイ	趣味とスポーツ	2
フランス	-	世界の国と地域	3

生成したカテゴリは，3 階層構造を成しており，カテゴリ数は約 2340 個である．また，各カテゴリにはそれぞれ，Web ページが分類してあるものとする．そして，

ユーザはカテゴリ誘導によって，カテゴリに登録されている Web ページを取得することができる．例えば,「ZZ-R」を検索したユーザに対しては，誘導されたカテゴリ「オートバイ」に登録されている Web ページを検索結果として出力する．

5.2 カテゴリ概念ベース (CCB)

カテゴリ誘導を行うための知識として，カテゴリ概念ベースを構築する．カテゴリ概念ベースは，カテゴリ名を概念として，その属性と重みの対の集合から構成されている．あるカテゴリ C を以下のように定義する．

$$\text{カテゴリ } C = \{(a_1, w_1), (a_2, w_2), \cdots, (a_m, w_m)\} \quad (5)$$

a_i は属性，w_i は重み，m はカテゴリ C の属性の総数である．カテゴリ総数約 2340 個に対して，属性と重みの付加を機械的に行う手法について以下で述べる．

5.3 Web を用いたカテゴリ概念ベースの構築

カテゴリは数も多く，固有名詞から成るものも多く存在するため，人手による属性と重みの付加は困難である．また，Web 空間上には膨大なデータが存在し，意味のある語について検索を行えば，ほとんどの場合，入力語に関する情報を取得することができる．そこで，Web を用いてカテゴリ概念ベースを機械的に構築する．構築手法を以下に示す．

まず，カテゴリ名をロボット型検索エンジンを用いて検索を行い，検索結果の上位 50 件の文書を取得する．そして，取得した文書に対して形態素解析を行い，名詞と動詞を取得する．取得した名詞・動詞をカテゴリ（概念）の属性として，重み付けを行い，カテゴリ概念ベースに格納する．重み付け手法としては，頻度（ＴＦ）のみによる重みとＴＦ・ＩＤＦを用いた重みの２種類を求め，比較を行った．自動構築したカテゴリ概念ベースの一部を表 2 に示す．

表 2: カテゴリ概念ベース

カテゴリ	属性・重み
肝炎	{(肝炎,360),(ウィルス,334), (感染,307),(検査,171),…}
サッカーくじ	{(予想,143),(くじ,141), (試合,43),(プレゼント,34),…}
自動車	{(新車,123),(一覧,80), (試乗,73),(保険,54),…}
教育	{(教育,140),(研究,38), (学校,32),(学習,21),…}

5.4 検索質問文

検索質問文によってユーザは自身の検索要求を具現化し，検索エンジンに入力する．そこで本稿ではある検索質問文 $Query$ をキーワード k_i と重み w_i の対の集合で以下のように定義する．

$$Query = \{(k_1, w_1), (k_2, w_2), \cdots, (k_L, w_L)\} \quad (6)$$

L は，ユーザが入力したキーワードの数である．

しかし，ユーザが検索エンジンに対して入力するキーワードは，1 語や 2 語と非常に少ない．そこで検索質問文に対しても，カテゴリ概念ベースの構築と同様に Web を用いて属性と重みを取得し，検索質問文の拡張を行った．

5.5 ベクトル空間モデルを用いたカテゴリ誘導

カテゴリ誘導を行うために，検索質問文と各カテゴリとの関連の深さを計算する．本稿では，検索質問文 $Query$ と各カテゴリ C 間の関連の深さを計算するのにベクトル空間モデルを用いた．$Query$ と C 間の関連の度合いを $Q-C$ 関連度とする．検索質問文 Q_i とカテゴリ C_j の $Q-C$ 関連度の値 $sim(Q_i, C_j)$ を以下の式で定義する．

$$sim(Q_i, C_j) = \cos\theta = \frac{Q_i \cdot C_j}{|Q_i||C_j|} \quad (7)$$

$sim(Q_i, C_j)$ の値が大きいほど，検索質問文とカテゴリとの関係が深くなる．

検索質問文とカテゴリ 2340 個すべてに対して，この $Q-C$ 関連度を求めることにより，検索質問文とカテゴリの関連の深さを定量化することができ，関連の深いカテゴリへと誘導することができる．

6. 実験と評価

属性数，重み付け手法別に CCB1〜CCB4 の 4 つのカテゴリ概念ベースの構築を行った．表 3 に，今回構築を行ったカテゴリ概念ベースの構成についてまとめる．属性数は，属性を重みの大きい順に並べたときの，上位の属性の個数である．

表 3: カテゴリ概念ベースの構成

種類	属性数	重み付け手法	備考
CCB_1	128	TF	
CCB_2	512	TF	
CCB_3	512	TF・IDF	
CCB_4	512	TF・IDF	1 文字属性除去

そして，自動構築したカテゴリ概念ベースの評価を行うために，テストセット 154 組を人手で作成した．表 4 にテストセットの一部を示す．

テストセットは，検索質問文とそれが適合するカテゴリのペアから成る．$Q-C$ 関連度を用いて，検索質問文に適合するカテゴリが全カテゴリ (約 2340 個) 中，何位に出力できるかを求めた．そして，適合するカテゴリが

表 4: テストセット

検索質問文	適合するカテゴリ
B型肝炎	肝炎
toto	サッカーくじ
プルコギ	グルメ
ピタゴラス	数学者

上位 1 位，10 位，20 位以内にどの程度の確率で出力できるかを評価した．評価結果を図 1 に示す．

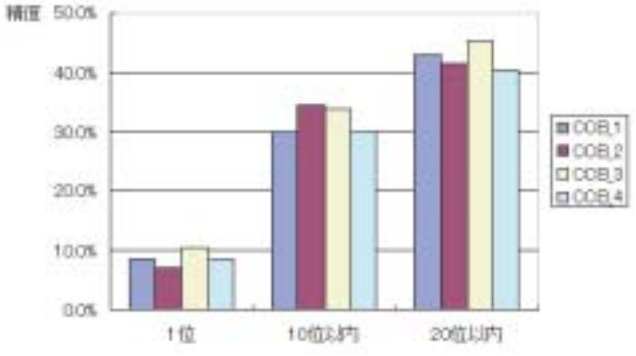

図 1: 適合するカテゴリの出現確率

図 1 より，適合するカテゴリの出現順位の平均が最もよいのは，カテゴリ概念ベース CCB_3 であることが分かる．この結果より，カテゴリ概念ベースの属性数はより多い方がよく，重み付け手法としてはＴＦよりもＴＦ・ＩＤＦを用いたほうがよいことが分かる．

CCB_4 で，1 文字属性を除去した理由は，形態素解析の問題で，1 文字の意味を持たない語がたくさん現れるため，1 文字属性を除去することにより，属性の精度が向上すると考えたからである．しかし，1 文字属性を除去することによって，精度が悪くなった．つまり，１文字属性には雑音も多く含まれているが，重みの大きい１文字のよい属性が多く存在していると考えることができる．

図 1 の結果は，人手で作成したテストセットによるものである．しかし，テストセットは人手で作成したために，検索質問文に対して適合するカテゴリを 100 %網羅できていない．例えば，テストセットでは，ユーザの検索質問文「B 型肝炎」に対して適合するカテゴリは「肝炎」であるが，カテゴリ「肝炎」以外にも適合するカテゴリが存在する．この場合,「肝炎」のほかに「病院，診療所」,「病気，症状」なども適合するカテゴリとしてよいと考えることができる．このように，検索質問文に対して適合するカテゴリを，100 %網羅するのは難しいので，さらに別の評価手法を用いて評価を行った．

そこで，テストセットに用いた検索質問文に対して，$Q-C$ 関連度によって得られた上位 10 件のカテゴリ中に，検索質問文と関連のあるカテゴリがどの程度含まれているかを調べた．評価には，図 1 より最も結果のよかったカテゴリ概念ベース CCB_3 を用いた．結果を図 2 に示す．

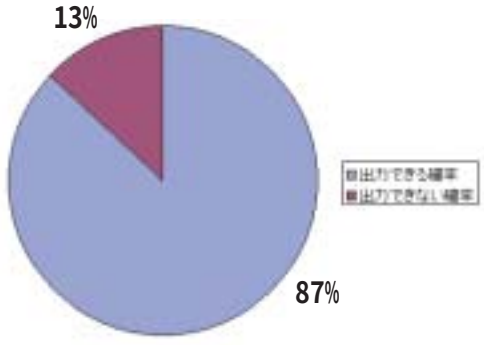

図 2: 関連のあるカテゴリを上位に出力できる確率

図 2 より，約 9 割の確率で検索質問文と関連のあるカテゴリが上位 10 位以内に出力できていることが分かった．

7. おわりに

本稿では，語と語の表記一致ではなく，語と語の意味の比較によるカテゴリ誘導手法について述べた．テストセットを用いた適合カテゴリの出現順位の平均は，20 位以内で約 45 %であった．しかし，上位 10 個の中のいずれかに適合するカテゴリが出現する確率は，約 90 %となった．つまり，$Q-C$ 関連度を用いることにより，カテゴリ 2340 個の中から，約 9 割の確率で上位 10 個以内に適合するカテゴリを出力することができることが分かった．

提案手法を用いることにより，固有名詞を含むどんな語からでも自動的に適切なカテゴリに誘導することができ，ユーザは効率よく欲しい情報を得ることができる．そして，ユーザの負担を軽減することができると考える．

本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクトにおける研究の一環として行った．

参考文献

[1] Salton,G.,Wong,A. and Yang,C.S.:”A vector space model for automatic indexing”,*Communications of the ACM*,Vol.18,No.11,pp.613-620,1975.

[2] Salton,G. and Buckley,C.:”Term-weighting approaches in automatic text retrieval”,*Information Processing & Management*,Vol.41,No.4,pp.513-523,1988.

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