3C4-02

Web検索における動作を観点とした情報分類方法

Operation based information classification for web information retrieval

堀田 真次　　渡部 広一　　河岡 司
Shinji Hotta　Hirokazu Watabe　Tsukasa Kawaoka

*1同志社大学工学研究科知識工学専攻
Department of Knowledge Engineering and Computer Sciences,Graduate School of Engineering,Doshisha University

This paper proposes the method of opareation based information classification. Verbs is acquired from Web using search engine and weight attachment is performed by TF・IDF which is the conventional method and LTF・COD which is the proposed method. Finally, this paper evaluates by experiment verb attributes and shows that Retrieval-Query with a viewpoint of operation is validity.

1. はじめに

近年，インターネット，World Wide Web（以下 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．この巨大な情報空間から，ユーザの検索要求に合った情報を探してくるのが WWW 検索エンジンである．

本研究では，ユーザの検索要求を明確にすることにより，情報をユーザの必要とするものとそうでないものに分類する．このように，検索対象とする情報をユーザの検索要求という条件によって，絞り込むことをフィルタリングと呼ぶ．フィルタリングを行うことにより，膨大な情報を整理し，ユーザの必要とする情報を見つけるための負担を軽減することが本研究の目的である．

2. 動作を観点とした情報分類

2.1 ユーザの検索要求の明確化

本研究では，カテゴリに分類されている Web ページを動作という別の観点を用いて，より詳細に分類する．

現在，情報を分類整理するカテゴリは名詞概念によって整理されている．しかし，カテゴリという名詞概念による分類では，ユーザの動作要求を考慮していないため，あいまいな分類になってしまう．そこで，カテゴリに対し，どのような動作要求があるかを推定する必要がある．カテゴリに対する動作要求をカテゴリ属性動詞とする．ユーザの動作要求を観点で分類することにより， ユーザが見る Web ページの量を減らすことができ，ユーザの負担を軽減することができる．

2.2 検索要求の推定と特定

ユーザの検索要求とは，情報検索を行う際にユーザが持っている要求である．検索質問文「BMW」は複数の検索要求を表現する.「BMW」に対する検索要求として,「BMW を買いたい」,「BMW を売りたい」などが挙げられる．

本研究では，ユーザの検索要求を明確にすることにより，よりユーザに適合する Web ページを提示することを目指す．検索要求を明確にすることを，ユーザが入力した名詞であるキーワードに対し「○○したい」という動作を追加することにより定義する.「○○したい」という動作を動作要求とする．

検索要求を明確にするために，ユーザの検索要求を推定する必要がある．そして，ユーザに推定した検索要求群を提示し，ユーザがその中から選択することにより，ユーザの動作要求を特定する．ユーザの動作要求を検索質問文に加えたものを，動作観点付検索質問文とする．例えば，検索質問文「BMW」に，ユーザの動作要求「購入する」を追加することにより，動作観点付検索質問文「BMW 購入する」を生成する．

連絡先: 堀田真次，同志社大学大学院工学研究科　知識情報処理研究室，〒 610-0394　京都府京田辺市多々羅都谷 1-3, 0774-65-6944

3. カテゴリ属性動詞の取得

カテゴリ属性動詞は，カテゴリに対する動作要求を表す動詞である．本稿では，カテゴリ属性動詞を Web から機械的に生成する．人手による生成も考えられるが，カテゴリ数が多く，取得には多くの労力を要するので機械的な処理が必要となる．カテゴリ属性動詞の取得の流れを以下に示す.

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し，その検索結果を解析する．そして，取得した文書に対して形態素解析を行い，動詞とサ変接続名詞を取得する．最後に，取得した動詞，サ変接続名詞に対して，以下で説明する TF・IDF[Salton 88]，LTF・COD それぞれについて重み付けを行い，それをカテゴリ属性動詞とする．取得したカテゴリ属性動詞の一例を表 1 に示す.

表 1: カテゴリ属性動詞

カテゴリ	カテゴリ属性動詞・重み
サッカーくじ	{(当選する,0.48),(購入する,0.16), ··· }
自動車	{(整備する,0.73),(入会する,0.51), ··· }

4. カテゴリ属性動詞の重み付け

Web から適切なカテゴリ属性動詞を取得するために，取得した動詞に重みを付与する．その結果カテゴリに対する適切なカテゴリ属性動詞をユーザに提示できる．本稿では，TF・IDF と LTF・COD という 2 つの重み付け手法を用いた.

4.1 ＴＦ・ＩＤＦを用いた重み付け

カテゴリ属性動詞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 による重み付けを行った．カテゴリ C_i に対するカテゴリ属性動詞 v_j の重み $w_i(v_j)$ を以下の式で定義する．

$$w_i(v_j) = tf_i(v_j) \cdot idf(v_j) \quad (1)$$

$$idf(v_j) = 1 + log_2(N/n_j) \quad (2)$$

$tf_i(v_j)$ はカテゴリ C_i に対するカテゴリ属性動詞 v_j の出現頻度，N は全カテゴリ数，n_j はカテゴリ属性動詞 v_j が出現するカテゴリ数である．

4.2 ＬＴＦ・ＣＯＤを用いた重み付け

$tf_i(v_j)$ は，最大値約 2000，最小値 1 であった．そこで，頻度の大きいカテゴリ属性動詞の重みの影響を小さくするために，対数化頻度を用いた．カテゴリ C_i に対する対数化頻度 $ltf_i(v_j)$ を以下のように定義する．

$$ltf_i(v_j) = log_2(1 + tf_i(v_j)) \quad (3)$$

また，式 2 の $idf(v_j)$ は，特定のカテゴリに出現する動詞ほど，重要な動詞であるという定義により重み付けを行っている．この重み付け手法では，専門的な動詞に対して大きな重みが与えられる．つまり，多くのカテゴリに出現する「購入する」,「売る」といった，情報検索をする際によく用いられる一般的な動詞の重みが小さくなってしまう．そこで，本稿では大局的重み付けとして共起結合度による重み付け手法を提案する．共起結合度は,「購入する」という動詞の分布だけではなく，動詞を付加する対象であるカテゴリを利用する．カテゴリ C_i に対する動詞 v_j の共起結合度 $cod_i(v_j)$ の式を以下のように定義する．

$$cod_i(v_j) = N_{ij}/NN_j \quad (4)$$

N_{ij} はカテゴリ C_i 中の文書の中で動詞 v_j が出現する文書数，NN_j は動詞 v_j が出現する文書数である．対数化頻度と共起結合度を用いて，カテゴリ C_i に対する動詞 v_j の重み $w_i(v_j)$ を以下の式で定義する．

$$w_i(v_j) = ltf_i(v_j) \cdot cod_i(v_j) \quad (5)$$

5. 実験と評価

カテゴリ属性動詞の TF・IDF，LTF・COD による重み付け手法の評価を行った．カテゴリを 50 個用意し，それぞれのカテゴリに対し，カテゴリ属性動詞を重みの大きい順に 5 個，10 個，20 個出力する．その内人手でどれくらい適切なカテゴリ属性動詞が存在するか評価する．50 カテゴリにおける平均適合率を図 1 に示す．

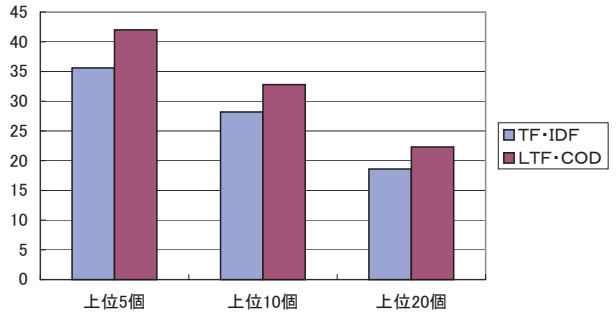

図 1: 属性動詞の評価

LTF・COD で重み付けを行ったカテゴリ属性動詞の上位 5 個出力したときの適合率は 42 %，上位 10 個の場合は 32.8 %，上位 20 個の場合は 22.3 %となった．また TF・IDF による重み付け手法を用いた場合と比べてすべての場合において適合率が向上することが分かる．ゆえに，本稿が提案した LTF・COD による重み付け手法は，カテゴリ属性動詞の重み付け手法として有効な重み付け手法であることが分かる．

6. 動作観点付検索質問文の有効性

入力キーワードに，ユーザのカテゴリ属性動詞の中から選択したユーザの動作要求を追加することにより，動作観点付検索質問文を生成した．動作観点付検索質問文の有効性を調べるために，次のような評価を行った．Google[Google] を用いて，ユーザの検索質問文のみで検索を行った結果と動作観点付検索質問文を用いて検索を行った結果の上位 10 件の適合率を調べた．評価結果を図 2 に示す．評価セットとして，10 セット用意し，平均適合率を求めた．評価セットには,「自動車 購入」,「サッカーくじ 予想」などがある．

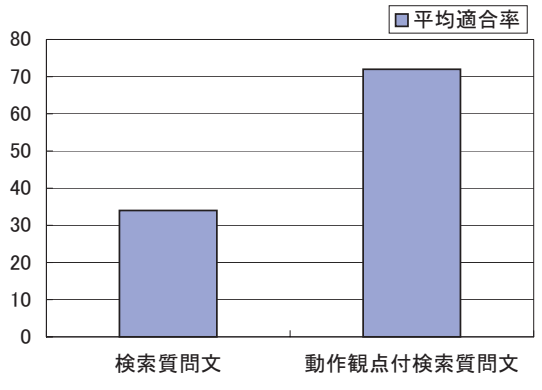

図 2: 動作観点付検索質問文の有効性

図 2 より，動作観点付検索質問文を用いることにより，検索結果上位 10 件の平均適合率は，38 %向上していることが分かる．つまり，動作観点付検索質問文を用いることにより，ユーザの検索要求に合った Web ページが上位に提示できた．ユーザの検索要求に適合する Web ページが上位に提示されるので，ユーザの見る Web ページの量を減らすことができ，ユーザの負担は軽減される．ゆえに，動作観点付検索質問文が有効であることがいえる．

7. おわりに

本稿では，カテゴリ属性動詞の重み付け手法の工夫を行った．本稿で提案する対数化頻度と共起結合度を用いた LTF・COD による重み付け手法が，TF・IDF による重み付け手法よりも有効であることが分かった．また，動作観点付検索質問文を用いることにより，より検索要求に適合する Web ページをユーザに提示できた．

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「知能情報科学とその応用」における研究の一環として行った．

参考文献

[Salton 88] Salton,G. and Buckley,C.:"Term-weighting approaches in automatic text retrieval",*Information Processing & Management*,Vol.41,No.4,pp.513-523,1988.

[Google] http://www.google.co.jp