専門性の程度に基づく Web 検索結果の分類手法

鶴田雅信, 梅村祥之, 酒井浩之, 増山繁

豊橋技術科学大学 知識情報工学系

1 はじめに

WEB 上に存在するさまざまな文書を, 一般の利用者が有効利用するには検索・分類技術 [1] が不可欠である. 従来の分類手法の多くは, 文書をジャンル別に分類するという考えのもとに考案されてきた [3]. 専門性の程度に関しては特に考慮の対象とされておらず, そのため, そのような手法の分類結果には, 専門知識を持っている利用者にとっては簡単すぎる, もしくは初心者にとっては難しすぎる文書が混在していた.

この問題を解決するためには, 利用者の知識に応じた文書を選択できるような分類を行うことが有効であると思われる. そこで, ある文書の専門性を指標化し, それによって WEB の検索結果を分類する手法を考案する. この手法では, 従来の手法とは異なり, 文書の内容の類似関係を使用しない. そのため, この手法によって文書をジャンル別に分類することはできない.

2 提案手法

2.1 専門性の定義

本研究では,「ある文書の専門性とは, その文書を読む際に, それの属する最も適当な分野に関する専門知識を, どのくらい要するか」ということとした.

2.2 専門性に関する文書の特徴

専門性の定義に従って人手による分類を行った. その結果に基づき「専門的な文書」と「専門的でない文書」の特徴を調査する.

専門的な文書の多くには, 以下のような特徴が見受けられる.

- 固い表現で書かれている.
- 略語やジャーゴン, 専門用語といった「重要語」が多数出現する.
- 含まれている「情報」の量が多い.

一方, 専門的でない文書には, 以下のような特徴が見受けられる.

- わかりやすい表現で書かれている.
- 一般的な語が使用される.
- 含まれている「情報」の量が少ない.

なお, ここでの「情報」の量とは, 数学的に定義された情報量のことではなく, あくまでも直観的な印象としての「情報」の量を意味する.

2.3 重要語とその出現密度

専門的な文書には, 重要語が多数出現する. 一方, 専門的ではない文書には, 専門的である文書ほど多くの重要語は出現しない. これらの印象より, ある文書の重要語の出現密度をその文書の専門性の指標として利用することで, 文書の専門性による分類を行うことができるのではないかと考えた.

2.3.1 重要語の定義

重要語とは, 検索結果に含まれる文書以外にはあまり出現しないが, 検索結果に含まれる文書中には多く出現する複合名詞 (2 語以上の名詞, もしくはカタカナ語が連結した語) のこととする.

2.4 手法

1. 検索結果である文書集合 S に含まれる複合名詞 $t_i, i = 1, 2, ..., n$ を抽出する.
2. t_i に対して, 語の重み $W(t_i, S)$ を計算する.

$$W(t_i, S) = (A + \frac{Tf(t_i, S)}{\max_i Tf(t_i, S)}) \times \log \frac{|N|}{df(t_i, N)} \times (B + \frac{En(t_i, S)}{\max_i En(t_i, S)}) \quad (1)$$

A, B：定数[*1]

[*1] 試行錯誤により $A = 0.4, B = 0.4$ とした.

$Tf(t_i, S)$：文書集合 S における語 t_i の出現頻度

$$Tf(t_i, S) = \sum_{s \in S} tf(t_i, s) \tag{2}$$

$tf(t_i, s)$：文書 s における語 t_i の出現頻度
$En(t_i, S)$：文書集合 S における語 t_i の出現確率に基づくエントロピー

$$En(t_i, S) = -\sum_{s \in S} P(t_i, s) \log_2(P(t_i, s)) \tag{3}$$

$$P(t_i, s) = \frac{tf(t_i, s)}{Tf(t_i, S)} \tag{4}$$

$df(t_i, N)$：全文書集合 N において, 語 t_i を含んでいる文書の頻度.

3. 重み W の上位 m 個を重要語として採用する.
4. 重要語の密度 $Den(s)$ を求める [2].

$$Den(s) = \frac{\sum_{t \in KS(s)} W(t, S)}{d(s)} \tag{5}$$

$$d(s) = \frac{\sqrt{\sum_{k=2}^{|KS(s)|} (dist_k)^2}}{|KS(s)| - 1} \tag{6}$$

$KS(s)$： s に出現する重要語の集合.
$dist_k$： k 番目に出現した重要語と $k-1$ 番目に出現した重要語の距離 (語の数).

5. 閾値を T としたとき,$Den(s) > T$ であれば文書 s は「専門的である」, そうでなければ「専門的ではない」と分類する. □

2.4.1 語の重み付け

(1) 式によって,「検索結果中には多く出現するが, 他の文書集合にはあまり出現しない」語に対し, 大きな重みの値が割り当てられる.

第一項では, 検索結果中の全文書における正規化された語の頻度を求めている. 検索結果中に最も多く出現する語において, この項は最も大きな値となる. 第二項は IDF 値であり, それぞれの文書において 1 度以上出現する確率が最も高い語のときに, 最も小さな値となる.

第三項では, 正規化された語のエントロピーを計算する. この項は, 1 文書にしか出現しない語のときに, 最も小さな値となる. 前二項の積を語の重みとしてそのまま使用すると, 1 文書にしか出現しないが, その 1 文書においては非常に多く出現するという語に対して, 高い重みを割り当ててしまう. そのような語は, 文書集合を代表する語という意味での「重要語」としては不適当である. よって, この項によってそのような語の重みを減らす必要がある. これらの三項を組み合わせることで, 重要語に対して大きな重みの値を割り当てることができる.

2.4.2 重要語の密度

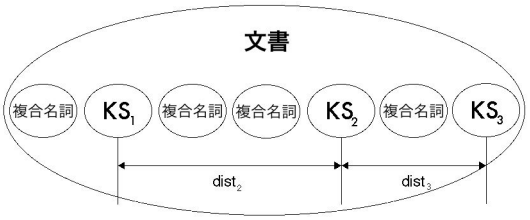

図 1　重要語密度の定義

重要語密度 $Den(s)$ を計算する式として, (5) 式を用いる [2] . この式には, (6) 式で示した, s に出現するキーワード間の 2 乗平均距離を使用している. 図 1 の例では, (6) 式における $|KW(S_i)| = 3$, $dist_2 = 2$, $dist_3 = 1$ である. この 2 乗平均距離が小さいということは, キーワードが文書において密集して出現することを意味する.

WEB 上に存在する文書には, 一文書に複数のテーマに関する記事が混在しているものがある. そのような文書の重要語密度を求める際に, 単純な重要語間の距離を使用すると, 文書の一部にのみ専門的かつ重要な内容を含む文書を, 「専門的でない」と判断してしまう可能性がある. 2 乗平均距離による重要語密度を利用する事で, そのような文書を, 効果的に分類することができると考えた.

3 実験

実験にあたっては, NTCIR ワークショップ [*2] の WEB タスク, トピック分類タスクのドライランに使用したターゲットデータセット (検索結果である文書集合) を利用した.

その内トピック番号 0028 を使用して実験を行った. このターゲットデータセットは, オーガナイザ側がベースライン検索システムにおいて「著作権, デジタルコンテンツ, ネットワーク」という検索クエリによって, 検索対象である文書集合中を検索した結果である. 今回はターゲットデータセットの上位 200 文書に対して分類処理を行った.

形態素解析には Mecab [*3]を使用した. また, 重要語の個数は全ての複合語のうち, 上位 100000 個とした.

ある複合名詞の中に別の複合名詞が含まれている場合, 最も長い複合名詞のみを使用した.

3.1 評価方法

分類対象となる検索結果中の文書に対し, 理工系学生 3 名がそれぞれ人手で「専門的である」, もしくは「専門的でない」とラベル付けする. そして,3 名による分類データの多数決を取り, その結果を正解データとする. その正解データと本手法による出力を比較し, 適合率, 再現率を算出する. なお, ここでの適合率と再現率とは, 以下の値とする.

$$\text{適合率} = \frac{\text{本手法により「専門的である」と分類された文書中の正解数}}{\text{本手法により「専門的である」と分類された文書数}}$$

$$\text{再現率} = \frac{\text{本手法により「専門的である」と分類された文書中の正解数}}{\text{正解データ中の「専門的である」文書数}}$$

3.2 閾値について

本手法で分類を行うためには, 閾値を決定しなければならない. そこで, 閾値を決定するための予備実験を行った.

[*2] http://research.nii.ac.jp/ntcweb/

[*3] http://cl.aist-nara.ac.jp/%7Etaku-ku/software/mecab/

3.2.1 閾値決定のための予備実験

ある検索結果の上位 200 文書を理工系学生 3 名が人手で分類し, その多数決を参考データとする. その参考データ中の「専門的である」とラベル付けされた文書と, 「専門的でない」とラベル付けされた文書の割合より閾値を決定する.

3.2.2 予備実験の結果

参考データにおいて,「専門的である」とラベル付けされた文書は 91 文書, 「専門的でない」とラベル付けされた文書は 109 文書であった. この結果より今回の実験では, 中央値を閾値として分類することにした.

3.3 実験結果

表 1 に人手の評価による再現率を, 表 2 , 図 2 に本実験の結果を示す. 人手の評価による再現率とは, 以下の値とする.

$$\text{人手の評価による再現率} = \frac{\text{ある評定者, および多数決によって「専門的である」とラベル付けされた文書数}}{\text{多数決によって「専門的である」とラベル付けされた文書数}}$$

表 1　人手の評価による再現率

評定者 1	0.94
評定者 2	0.69
評定者 3	0.83

表 2　実験結果

再現率	0.68
適合率	0.68

正しく分類できた文書の特徴は：

- 用語集, 書籍発売データなどが専門的であると分類.
- 初心者向けの短い解説文などが専門的でないと分類.

分類に失敗した文書の特徴は：

- 日記, 掲示板などが専門的であると分類.
- 画像主体の文書などが専門的でないと分類.

出版社: 9.52454543319945
特開 : 4.29863132089488
著作物 : 3.06737976089705
デジタルコンテンツビジネス : 2.98331473688866
デジタルコンテンツ : 2.94948732700488
著作権 : 2.75468239298914
マルチメディアニュース : 2.73735132560551
東北デジタル放送 : 2.52322809925965
電子透かし : 2.50765895341282
コンテンツ主体 : 2.4142784111529
穂波町 : 2.3861892343983
出版コンテンツ : 2.32456911739937
電子商取引 : 2.30926893880369
サテライトニュース : 2.23202051142928
著作権保護 : 2.23058036846664
音楽産業 : 2.228790285458
電子商取引環境整備 : 2.22844507608736
著作権制度 : 2.22241895534056
情報処理装置 : 2.2092527617562
衛星関連ニュース : 2.19776079521394

図 2　重要語重み上位 20 個

であった.

4　考察

人手の評価による再現率の平均が 0.82 であったので, 上記の実験結果は良好であったといえる. 重要語の密度は専門性の指標として, ある程度有効であることが確認できた.

分類に失敗した文書として, 日記, 掲示板などといったものがあった. そのような文書には, その場所のみで多用される固有名詞, ジャーゴンなどが存在することがある. 今回の手法で使用した語の重みの計算式は, それらの影響を受けにくくなっている. しかし, 同じ掲示板の複数のログファイルが検索結果に含まれている場合などには対処できなかった. また, 例えば日記であれば同じ日付, 掲示板であれば同じ投稿記事に, それらの固有名詞, ジャーゴンが密集しているという特徴があった. この場合は, 語の出現密度の計算式が裏目に出たという結果になった.

語の重みは, 意図通りの結果となった.

重要語の密度がゼロであった文書も存在した. これらは文書が短すぎるといった原因のため, 複合名詞, もしくは重要語が 2 つ以上存在しなかったものであった.

5　今後の課題

検索結果内の文書の割合は, 検索語によって大きく変わってくる可能性がある. その場合, その割合を利用者が設定できれば, より利用しやすいシステムになる. また, 文体の情報を併用することで, さらに精度が向上する可能性がある.

重要後の密度がゼロになる問題は, 重要語として複合名詞のみを使用したためである. これを避けるためには, 複合名詞以外から有効な重要語を抽出しなければならない.

また,2 語以上の複合名詞について, 単純に語を繋ぎ合わせるのではなく, もっとも有効な組み合わせを統計的手法などによって探し出すことで, より有効な重要語密度を算出することができると考えられる.

謝辞

本研究は文部科学省２１世紀 COE プログラム「インテリジェント ヒューマンセンシング」および日本学術振興会科学研究費基盤研究 (C)(2)1368044 の援助により行なわれた.

参考文献

[1] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, Addison Wesley, 1999.

[2] H. P. Luhn, The Automatic Creation of Literature Abstracts, I. Mani and M. T. Maybury eds., Advances in Automatic Text Summarization, The MIT Press, 1999. (originally presented at IRE National Convention, 1958)

[3] C. D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, The MIT Press, 1999.