局所的IDFを用いた検索結果の再ランキング手法

平川雄三† 鈴木優† 川越恭二†
†立命館大学 情報理工学部

1 はじめに

近年，検索エンジンを利用して情報検索を行う機会が増加している．ところが，現在の検索エンジンは問合せを含む文書を検索対象としており，検索結果における文書は必ずしも問合せに適合した内容であるとは限らない．このような場合に，TF-IDF 法を用いることによって，問合せに対する重要度順に検索結果を再ランキングする解決手法が考えられる．ただし，従来の TF-IDF 法では，各文書に対して同一の単語に同一の IDF を付与している．このため，文書の内容が異なる場合に，文書を最適にランキングすることが困難であると考えられる．例えば，沖縄の海に関する文書と沖縄の水族館に関する文書がある場合，沖縄の水族館に関する文書に含まれる「海」という単語よりも，沖縄の海に関する文書に含まれる「海」という単語が評価されるべきであると考えられる．

そこで本研究では，単語の重要度は文書の内容によって異なる点に着目し，文書の内容を考慮した単語の重要度を用いることによって，検索結果を再ランキングする手法の提案を行う．まず，提案手法は分野別にあらかじめ分類された文書集合と文書との文書間類似度によって，文書が属する分野を決定する．次に，ある単語が文書の属する分野にどれだけ集中して出現しているかを表す値を算出する．この値が高い単語は，その文書において重要な単語であると考えられる．算出した値を用いて文書に重要度を付与することによって，分野を考慮した重要度順に検索結果の文書を再ランキングする．このことにより，検索結果において異なる分野に属する文書が混在する場合に，利用者の問合せに適合した文書の検索精度が向上すると考えられる．

2 関連研究

村松ら [1] は Web 全体集合と問合せに対する検索結果集合において，各単語の特徴量の違いを考慮することによって，検索結果中の単語の専門性，一般的認知度にもとづく分類を行い，検索結果にラベリングを行っている．本研究は，検索結果の各文書を分野別の文書集合に対応付け，検索結果の各文書における単語の特徴量を求めている点において村松らの手法と異なる．

Reranking Method of Search Results using Local IDF
Yuzo HIRAKAWA†, Yu SUZUKI†and Kyoji KAWAGOE†
†College of Information Science and Engineering, Ritsumeikan University.
{hirakawa, suzuki, kawagoe}@coms.ics.ritsumei.ac.jp

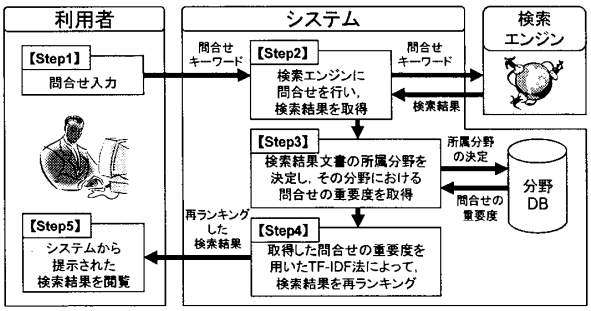

図 1: 提案手法の概要

3 局所的IDFを用いた検索結果の再ランキング手法

本研究では，文書が属する分野を Yahoo!カテゴリ[1] が提供している各カテゴリであると定義する．ここで分野とは，文書における内容別の分類であり，例えば，「料理」や「携帯電話」といった分野が挙げられる．Yahoo!カテゴリは人手によって文書が内容別に分類されているため，各カテゴリが分野に相当すると考えられる．

3.1 提案手法の概要

提案手法の概要を図 1 に示す．まず Step1 として，利用者はシステムに問合せを入力する．次に Step2 において，システムは利用者が入力した問合せに対する検索結果を検索エンジンから取得する．Step3 において，システムは分野情報のデータベースを利用することによって，検索結果の各文書と各分野に含まれる文書集合との類似度を算出する．算出した類似度を用いて，検索結果の各文書が属する分野を決定する．そして，システムは文書ごとに問合せの重要度を取得する．Step4 において，システムは取得した重要度を用いた TF-IDF 法によって，分野を考慮した重要度順に検索結果の文書を再ランキングする．最後に Step5 において，利用者は再ランキングされた検索結果を閲覧する．

3.2 分野を考慮した単語の重要度の算出

局所的 IDF と大域的 IDF を用いることによって，分野を考慮した単語の重要度を算出する．本研究における局所的 IDF とは，あるカテゴリに含まれる文書集合における単語の IDF である．また，複数のカテゴリから文書を収集し，それらの文書集合における単語の

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

IDF を大域的 IDF とする．局所的 IDF と大域的 IDF を用いることによって，ある単語が特定の分野にどれだけ集中して出現しているかを表す値を算出することが可能であると考えられる．本研究では，局所的 IDF に対する大域的 IDF の比の値を文書が属する分野を考慮した単語の重要度とする．なぜならば，文書が属する分野に集中して出現している単語はその文書において重要度が高いと考えられるためである．

Yahoo!カテゴリから取得した M 個のカテゴリを $C_1, C_2, \cdots, C_M$ とし，カテゴリ $C_i (1 \leq i \leq M)$ にサブカテゴリ内の文書も含めて N 件の文書が含まれているとする．また，C_i から重複なく抽出された m 個の単語を $W_{i1}, W_{i2}, \cdots, W_{im}$ とし，単語 $W_{ij} (1 \leq j \leq m)$ の局所的 IDF を $g_{local_{ij}}$，大域的 IDF を g_{global_j} とする．C_i において W_{ij} が出現する文書数を d_{ij} として，$g_{local_{ij}}$ を算出する式を (1) 式に示す．

$$g_{local_{ij}} = \log\left(\frac{N+1}{d_{ij}}\right) \tag{1}$$

Yahoo!カテゴリから収集した全 N' 件の文書の集合において，W_{ij} が出現する文書数を d'_j とし，g_{global_j} を算出する式を (2) 式に示す．

$$g_{global_j} = \log\left(\frac{N'+1}{d'_j}\right) \tag{2}$$

$g_{local_{ij}}$ に対する g_{global_j} の比の値 g_{ij} を C_i における W_{ij} の重要度とする．g_{ij} を算出する式を (3) 式に示す．

$$g_{ij} = \frac{g_{global_j}}{g_{local_{ij}}} \tag{3}$$

g_{ij} が大きいほど，単語 W_{ij} はカテゴリ C_i において重要だと考えられる．

3.3 文書が属する分野の決定

検索結果の各文書の特徴ベクトルと各カテゴリの特徴ベクトルを生成し，特徴ベクトル間のコサイン尺度によって，検索結果の各文書が属する分野を決定する．検索エンジンから取得した n 件の検索結果文書を $D_1, D_2, \cdots, D_n$ とし，文書 $D_k (1 \leq k \leq n)$ が属するカテゴリ C_i を決定する．このために，D_k の特徴ベクトル $\boldsymbol{a}_k$ と C_i の特徴ベクトル $\boldsymbol{b}_i$ を生成する．特徴ベクトルの要素には，文書の内容を特徴付けるために有効であると考えられるため，TF-IDF による単語の重みを用いる．ただし，IDF には g_{global_j} を用いる．まず，D_k から抽出された単語の重みを $w_{k1}, w_{k2}, \cdots, w_{km}$ とする．ここで，C_i に含まれているが D_k には含まれていない単語の重みは 0 とする．そして，$w_{kp} (1 \leq p \leq m)$ を要素とした D_k の特徴ベクトル $\boldsymbol{a}_k = (w_{k1}, w_{k2}, \cdots, w_{km})$ を生成する．次に，C_i に含まれる W_{ij} の重みを $w'_{i1}, w'_{i2}, \cdots, w'_{im}$ とし，$w'_{ij} (1 \leq j \leq m)$ を要素とした C_i の特徴ベクトル $\boldsymbol{b}_i = (w'_{i1}, w'_{i2}, \cdots, w'_{im})$ を生成する．最後に，D_k と C_i 間の類似度を $\boldsymbol{a}_k$ と $\boldsymbol{b}_i$ のコサイン尺度によって算出する．コサイン尺度が大きいほど，検索結果の文書とカテゴリは類似していると考えられるため，D_k をコサイン尺度が最も大きいカテゴリ C_i に所属させる．

3.4 検索結果の再ランキング

利用者の問合せに対する検索結果の各文書を，各文書における単語の重要度を用いて再ランキングする．重要度の高い単語が多く出現している文書ほど重要であると考えられるため，従来の TF-IDF 法における IDF の代わりに g_{ij} を用いた手法によって，文書に重要度を付与する．文書 D_k がカテゴリ C_i に属する場合，問合せ $Q = \{q_1, \cdots, q_j, \cdots, q_r\}$ に対する D_k の重要度 $score(Q, D_k)$ を算出する式を (4) 式に示す．

$$score(Q, D_k) = \sum_{j=1}^{r}\left[\frac{l_{jk} g_{ij}}{0.8 + 0.2\dfrac{t_k}{t_{avg}}}\right] \tag{4}$$

ここで，l_{jk} は文書 D_k における q_j の出現個数，g_{ij} はカテゴリ C_i に属する文書 D_k における q_j の重要度である．q_j が C_i 内に出現していない場合は g_{ij} を 0 とする．また，$0.8 + 0.2\frac{t_k}{t_{avg}}$ は文書正規化係数，t_k は文書 D_k に含まれる単語数，t_{avg} は検索結果文書に含まれる平均単語数である．$score(Q, D_k)$ にもとづいて，利用者の問合せに対する重要度が高い順に検索結果の文書を再ランキングする．

4 おわりに

本研究では，単語の重要度は文書の内容によって異なる点に着目し，文書の内容が属する分野ごとに算出した単語の重要度を用いることによって，検索結果を再ランキングする手法について提案した．検索結果において異なる分野に属する文書が混在する場合，提案手法による再ランキングを行うことによって，利用者の問合せに適合した文書の検索精度が向上すると考えられる．今後は，提案手法の有効性を検証するための実験を行い，実験結果について考察する．また本研究では，特定の分野に集中して出現している単語を重要単語であると考えたが，村松ら [1] や高見ら [2] の研究のように，多角的な視点で単語に重要度を付与することも検討する予定である．このことにより，単語の重要度が文書の内容によって異なることを考慮し，様々な検索目的に対応することが可能になると考えられる．

参考文献

[1] 村松亮介, 横山昌平, 福田直樹, 石川博：“単語の特徴量を考慮した検索結果クラスタに関する多視点融合型スニペットの構築”, 情報処理学会研究報告. DBS, データベースシステム, **146**, 51, pp. 301–306 (2008).

[2] 高見真也, 田中克己：“検索目的に基づくスニペットの動的再生成によるウェブ検索結果の個人適応化”, 日本データベース学会 Letters, **6**, 2, pp. 33–36 (2007).