5U-04 非適合文書プロファイルを利用した文書フィルタリング手法の検討

大西 亜希子 † 帆足 啓一郎 ‡ 橋本 和夫 ‡ 白井 克彦 †

†早稲田大学理工学部 ‡KDD 研究所

1 はじめに

文書フィルタリングを行う際，精度を向上させるために，情報検索における検索式拡張手法をプロファイル更新に適用する手法の有効性が報告されている [1] が，依然として多くの非適合文書を誤って選択する結果となっている．本研究では，文書フィルタリング中に選択される非適合文書を減少させるため，過去に選択された非適合文書の特徴を表す非適合文書プロファイルを作成し，二つ目のフィルタとして使用することで精度向上を図る手法を提案する．

2 単語寄与度を利用したプロファイル更新手法

フィルタリングにおけるプロファイル更新には情報検索における検索式拡張手法を適用する手法が多く用いられている．ここでは，検索式拡張手法の一つである，単語寄与度を利用した検索式拡張手法 [2] をプロファイル更新に適用した手法について説明を行う．

単語寄与度とは，文書間の類似度における各単語の影響を数値化した尺度である．参考文献 [2] で述べられた検索式拡張手法では，適合文書に出現する単語のうち，単語寄与度が大きく負の値を持つ単語を検索式拡張に利用した．

ここではフィルタリングにおいて，個々の検索対象文書 d に対して以下のアルゴリズムによる処理を行うことで，プロファイル更新を行う．ただし，プロファイル q_R を表すベクトルを $\vec{q_R} = (q_{R_1}, q_{R_2}, \cdots, q_{R_n})$ とし，$Sim(q_R, d)$ はプロファイル q_R と文書 d との間の類似度を，$Cont(w_i, q_R, d)$ はプロファイル q_R と文書 d との間の類似度における単語 w_i の単語寄与度を，$idf(w_i)$ は単語 w_i の Inversed Document Frequency を表す．また，パラメータとして wgt_{rel_R}，wgt_{nrel_R} を使用する．

プロファイル更新アルゴリズム

```
if(Sim(q_R, d) > 閾値){      /* 以下，プロファイル更新 */
  単語寄与度に基づいて単語を抽出
  if(d ∈ 適合文書){
    /* (1) 適合文書の場合 */
    foreach(抽出された単語 w_i){
      Score(w_i) = wgt_{rel_R} × Cont(w_i, q_R, d) × idf(w_i)
      q_{R_i} = q_{R_i} + Score(w_i)
    }
  } else {
    /* (2) 非適合文書の場合 */
    foreach(抽出された単語 w_i){
      Score(w_i) = wgt_{nrel_R} × Cont(w_i, q_R, d) × idf(w_i)
      q_{R_i} = q_{R_i} − Score(w_i)
    }
  }
}
```

本手法においては，上記のように更新されるプロファイルと検索対象文書との類似度を計算し，類似度が閾値を超えた場合その文書を選択する．しかし，この手法ではプロファイルと適合文書および非適合文書の類似度が近い場合，適切な閾値を設定することが困難であり，多くの非適合文書が選択されてしまう可能性がある．

3 非適合文書プロファイルを利用したフィルタリング

誤って選択される非適合文書数を減らし，フィルタリングの精度向上を図るために，適合文書を表す q_R とは逆に，誤って選択された非適合文書の特徴を表現する“非適合文書プロファイル”(以下，q_N) を作成し，非適合文書プロファイルとの類似度が高い文書は選択しないという，新たなフィルタを導入する．

そして，従来通り適合文書を表す q_R との類似度がある閾値より大きいかどうかを調べるフィルタと，非適合文書を表す q_N との類似度が閾値より小さいかどうか

Document filtering method with non-relevant document profile.

Akiko Onishi †,Keiichiro Hoashi ‡, Kazuo Hashimoto ‡, and Katsuhiko Shirai †.

†School of Science and Engineering,Waseda University.

‡KDD R&D Laboratories.

を調べるフィルタを使用した2重のフィルタリングを行う．これにより，q_Rに類似していると判断された文書から，過去に誤って選択された非適合文書に類似している文書を除外し，選択される非適合文書数を減少させることができると期待される．

q_Nは，2節で述べたプロファイル更新アルゴリズムにおいて，閾値を超えた文書が適合文書の場合には(2)の処理を，非適合文書の場合には(1)の処理を行うことにより作成する．その際，適合文書・非適合文書それぞれにq_Rとは異なるパラメータwgt_{rel_N}，wgt_{nrel_N}を用いる．

4 評価実験

非適合文書プロファイルを利用したフィルタリング手法の評価を行うために以下の実験を行った．

手法

実験には，TREC8 の Filtering Track において与えられる，時間順に整列した約20万件の検索対象文書と，50件のプロファイル，各プロファイルの適合文書のセットを使用した．また，フィルタリングの評価に，情報検索で用いられる Recall,Precision を用いることは適切ではないため，本研究では TREC で用いられている *Scaled Utility*[1] を使用して評価を行った．

q_Rとの類似度の閾値 $Thres_R$を 0.1，$wgt_{rel_R} = -200$，$wgt_{nrel_R} = -800$ とし，提案手法においてはq_Nとの類似度の閾値$Thres_N$を0.25とし，wgt_{rel_N}={-200,-400,-800}，wgt_{nrel_N}={-100,-200,-400,-800} の各値において実験を行った．

結果

図1に，$wgt_{rel_N} = -800$，$wgt_{nrel_N} = -200$ においてq_Rとの類似度が$Thres_R$を超えた文書の，q_Nとの類似度 sim_Rと q_Nとの類似度 sim_Nとの関係を適合文書・非適合文書で区別して示す．また，wgt_{rel_N}，wgt_{nrel_N}の各組み合わせにおける全プロファイルの平均 *Scaled Utility*を表1に示す．

表 1: 平均 *Scaled Utility*

	wgt_{nrel_N}			
wgt_{rel_N}	-100	-200	-400	-800
-200	0.5448	0.5464	0.5448	0.5508
-400	0.5448	0.5466	0.5491	0.5466
-800	0.5408	0.5466	0.5484	0.5505
q_Rのみ	0.5257			

図1より，$Thres_R$を超える sim_Rにおいては適合文書と非適合文書が混在しており，多くの適合文書を取得するために閾値を低く設定すると多くの非適合文書を選択してしまい，非適合文書の選択数を減少させるために閾値を高く設定すると適合文書の取得数も減少してしまうことが分かる．

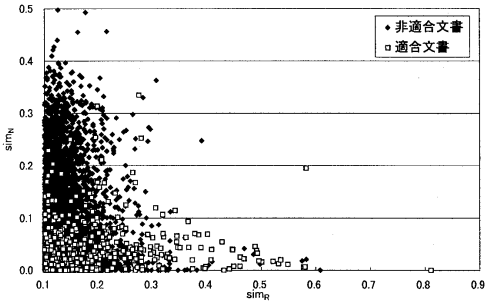

図 1: 各文書と q_R，q_Nとの類似度

一方でsim_Nに関しては，適合文書は全体にsim_Nが小さいものが多く，非適合文書はsim_Nの高い位置にまで分布していることが分かる．このことから，sim_Nに閾値を設定することで，q_Rとの類似度のみを基準としたフィルタリングで誤って選択される非適合文書を減らし，より精度の高いフィルタリングを実現することが出来る．

また，表1より，従来のフィルター一つを用いた手法と比較して各パラメータ設定において *Scaled Utility* が向上していることが分かる．

以上の結果から，非適合文書プロファイルを利用したフィルタリングの有効性が示された．

5 結論

非適合文書プロファイルを利用した手法の評価実験において，従来の適合文書を表すプロファイルとの類似度が閾値を超えたすべての文書に対し，非適合文書プロファイルとの類似度を測定したところ，非適合文書の類似度が適合文書と比較して高くなっていることが分かった．また，*Scaled Utility*には3〜5%の向上が見られた．以上より，非適合文書プロファイルを利用した手法の有効性が確認された．

参考文献

[1] D A Hull:"The TREC-7 Filtering Track : Description and Analysis" , The Seventh Text REtrieval Conference, pp33-56, 1999.

[2] 帆足，松本，井ノ上，橋本："文書間の類似度における単語寄与度を利用した検索式拡張手法"，情報処理学会論文誌：データベース，Vol.40 No.SIG8， pp63-73，1999.