隣接文書の特徴を考慮した文書特徴付け手法の精度比較

田村 航弥† 波多野 賢治‡ 宿久 洋‡
†同志社大学大学院文化情報学研究科 ‡同志社大学文化情報学部

1 はじめに

今日のインターネットの普及によって作られた大量なデータは，Web利用者(ユーザ)にとって必要な情報を得るための有用な情報源である．その反面，大量データの中からユーザにとって有用な情報を発見することが困難であるという問題も発生する．一般にインターネットを介して情報を得る際にはWeb検索エンジンを用いるが，Web検索エンジンがユーザにWeb文書を提示するためには，Web文書を何らかの形で数値に置き換える特徴付けの処理が必要である．この処理に対して，各Web文書の特徴を的確に捉え特徴付けすることで，ユーザの情報要求に適したWeb文書を，検索結果として提示することが可能であると考えられる．

一般的な文書特徴付け手法は，各文書に出現している索引語の統計量を算出し，それを各文書の特徴量として扱うTF-IDF法[3]やクエリ尤度モデル[4]などが提案されているが，過去の研究においてはこの方法に問題があるとして隣接文書の特徴量を付加して新たな特徴量を生成する改良手法が提案されている[1, 2]．このそれぞれの改良手法の検索精度において隣接文書を考慮しない文書特徴付け手法よりも良い成果を挙げている．しかし，一般的にはTF-IDF法を用いた文書検索よりもクエリ尤度モデルによる文書検索の方が検索精度が向上するという報告されているが[4]，隣接文書を考慮した場合のそれぞれの検索手法の精度比較は行われていない．

そこで本稿では，上述した二種類の特徴付け手法に対して，隣接文書の特徴量を考慮する改良手法の実装による検索精度の比較を行い，その考察を行う．

2 関連研究

2.1 隣接文書の内容を考慮したTF-IDF法の改良手法

杉山らは，リンク解析に基づくWeb検索では，文書間の内容の関連性を考慮できていないという問題点があると指摘している．それを受け，隣接文書の内容を考慮したTF-IDF法の改良手法の提案をしている[1]．この手法では，対象とするWeb文書の特徴ベクトル $p_{tgt} = (t_1, t_2, \ldots, t_n)$ における索引語 t_k のTF-IDF法に基づく重み $w_{t_k}^{p_{tgt}}$ に対して以下の式を用いて再計算している．

$$
\begin{aligned}
w_{t_k}^{'p_{tgt}} &= w_{t_k}^{p_{tgt}} \\
&+ \frac{1}{Dim}\left(\sum_{i=1}^{L_{(in)}}\sum_{j=1}^{N_{i_{(in)}}} \frac{w_{t_k}^{p_{ij_{(in)}}}}{N_{i_{(in)}} \cdot dis(p_{tgt}, p_{ij_{(in)}})}\right) \\
&+ \frac{1}{Dim}\left(\sum_{i=1}^{L_{(out)}}\sum_{j=1}^{N_{i_{(out)}}} \frac{w_{t_k}^{p_{ij_{(out)}}}}{N_{i_{(out)}} \cdot dis(p_{tgt}, p_{ij_{(out)}})}\right) \quad (1)
\end{aligned}
$$

ここで，$w_{t_k}^{'p_{tgt}}$ は対象とするWeb文書に対して L ホップのin-linkとout-linkで接続されたWeb文書の特徴ベクトル $p_{ij_{(in)}}$，$p_{ij_{(out)}}$ に含まれている索引語の重みで再計算された値である．dim は索引語の異なり語数を表し，$dis(p_{tgt}, p_{ij})$ は2文書間の距離を表しており以下のように定義される．

$$
dis(p_{tgt}, p_{ij}) = \sqrt{\sum_{k=1}^{m}(w_{t_k}^{p_{tgt}} - w_{t_k}^{p_{ij}})^2} \quad (2)
$$

このように新たに特徴づけされたWeb文書ベクトル p'_{tgt} とクエリとして与えれらた索引語をベクトルで表現したクエリベクトル $Q = (t_1, t_2, \ldots, t_n)$ との類似度 $sim(p'_{tgt}, Q)$ は以下の式によって計算する．

$$
sim(p'_{tgt}, Q) = \frac{p'_{tgt} \cdot Q}{|p'_{tgt}| \cdot |Q|} \quad (3)
$$

2.2 隣接文書の内容を考慮したクエリ尤度モデルの改良手法

近年の情報検索の研究分野においては，TF-IDF法を用いた文書特徴付け手法より高い検索精度を実現している言語モデルを用いた情報検索モデル[4]が多く用いられるようになった．この手法は，各文書における言語現象を確率的言語モデルを用いて推測し，クエリとして与えられた索引語が出現する確率を算出する検索モデルである．ここで文書 d_h で推測された言語現象を文書モデル M_{d_h} と呼び，与えられたクエリ Q が文書 d_h に出現する確率 $P(Q|M_{d_h})$ をクエリ尤度と呼ぶ．

この特徴付け手法に対して我々は過去の研究において，隣接文書のクエリ尤度を考慮した文書特徴付け手

Comparative Study of Characterizing Methods Considering Neighbor Pages
Kohya Tamura† ,Kenji Hatano‡ and Hiroshi Yadohisa‡
†Graduate School of Culture and Information Science, Doshisha University
‡Faculty of Culture and Information Science, Doshisha University

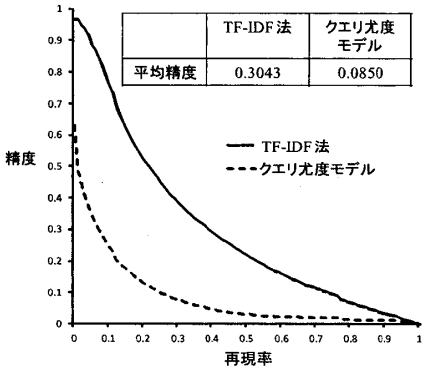
法を提案している [2]．この手法は，あらかじめ算出されたクエリ尤度に対して隣接文書のクエリ尤度を付与する方法であり，以下の式を用いて算出する．

$$P'(Q|M_{d_h}) = P(Q|M_{d_h})((\sum_{u_{h_l} \in U_{d_h}}^{s} P(Q|M_{u_{h_l}})) + 1) \quad (4)$$

ここで U_{d_h} は文書 d_h と隣接している文書集合であり，隣接文書 $u_{h_1}, u_{h_2}, \ldots, u_{h_s}$ から構成される．

3 評価実験

本実験は，上述した二つの手法について検索精度の比較を行うために INEX テストコレクション* を用いて実験を行った．本実験では評価尺度として 101 点平均精度及び再現率- 精度グラフを用いている．



	TF-IDF 法	クエリ尤度モデル
平均精度	0.3043	0.0850

図 1: 実験結果

実験の結果，杉山らの手法である隣接文書の内容を考慮した TF-IDF 法の改良手法による検索精度が我々の手法である隣接文書の内容を考慮したクエリ尤度モデルの改良手法による検索精度を大きく上回る結果となった (図 1)．このような結果となった原因は，比較した二つの手法における特徴量の再計算方法の差が挙げられる．杉山らの手法では，式 (1) より文書の特徴量を再計算する際に隣接する文書数によって付加する特徴量を正規化していることがわかる．これに対して我々の手法では，式 (4) より隣接文書数によって正規化がされていない．これによってたとえクエリと適合していない文書であっても，隣接文書数が多ければその文書は検索結果の上位にランキングされる．

この原因を裏付けるための分析として，クエリ尤度モデルでの検索精度と我々の手法での検索精度の比較を，検索課題として用いたクエリごとに行った．そして我々の手法が，クエリ尤度モデルの検索精度よりも下回った検索結果について，検索結果の上位 100 件の Web 文書に隣接している文書数の平均値を算出した．その結果，文書全体の平均隣接文書数は 37.67 文書であるのに対して上述した文書の平均隣接文書数は 198.52 文書であった．よって，検索精度の低下しているクエリの検索結果上位の文書は隣接文書を多く保持する傾向にある．

しかし，文献 [2] では隣接文書数の平均値によって付加する特徴量を正規化する手法を提案しているが，式 (4) の特徴付け手法用いたことによる検索結果での検索精度を下回る検索結果となっている．よって，隣接文書数の平均をもって正規化するのではなく，クエリと適合している隣接文書のみを抽出し特徴量の再計算に用いるという新たな手法を提案する必要がある．

4 まとめ

本稿は，Web 文書をより的確に特徴付ける手法である TF-IDF 法とクエリ尤度モデルによる隣接文書の内容を考慮した文書特徴付け手法の検索精度比較を行った．比較実験の結果，TF-IDF 法の改良手法の検索精度が上回る結果となり，クエリ尤度モデルの改良手法における問題点を発見することができた．今後の課題として，与えられたクエリとの適合度が低い隣接文書の扱い方の検討が挙げられる．

謝辞

本研究の一部は，独立行政法人日本学術振興会 科学研究費補助金 基盤研究 (C) (課題番号 : 21500284) によるものである．ここに記して謝意を表す．

参考文献

[1] 杉山一成, 波多野賢治, 吉川正俊, 植村俊亮. ハイパリンクで結ばれた隣接ページの内容に基づく Web ページのための TF-IDF 法の改良. 電子情報通信学会論文誌, Vol. J87-D-I, No. 2, pp. 113–125, 2004.

[2] 田村航弥, 波多野賢治, 宿久洋. 隣接ページのクエリ尤度を考慮した文書特徴付け手法の実装とその評価. 情報処理学会研究報告, 第 2009-DBS-149 巻, No, 3, pp. 1–8, 2009.

[3] G. Salton and C. Buckley. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, Vol. 24, No. 5, pp. 513–523, 1988.

[4] F. Song and W. B. Croft. A General Language Model for Information Retrieval. In *CIKM '99*, pp. 316–321. ACM, 1999.

*http://www.inex.otago.ac.nz/data/documentcollection.asp