The 29th Symposium on Information Theory
and its Applications (SITA2006)
Hakodate, Hokkaido, Japan, Nov. 28 – Dec. 1, 2006

単語ごとの修正重みに基づく適合性フィードバックによる文書検索

Document Retrieval by Relevance Feedback Based on Modified Weight of Each Word

平松 丈嗣* 石田 崇* 平澤 茂一*
Joji HIRAMATSU Takashi ISHIDA Shigeichi HIRASAWA

Abstract— The vector space model is a typical information retrieval model. The technique to improve the retrieval result using VSM uses the query expansion. One of the way of query expansion is relevance feedback by Rocchio algorithm. Another way of query expansion is to use of word contribution, which is numerical individual word that influence the accuracy of retrieval.

In this paper, to obtain the weight of the word which is specialized in the relevant documents, we focus upon the appearance frequency of the word in the relevant documents. The weight of the word will be large when the appearance frequency is high. Obtained weights are is added to the first query vector, and retrieved again by using a new query vector. As a result, our method of the accuracy of retrieval is enhanced compared to that of the conventional methods.

Keywords— Vector Space Model, Relevance Feedback, Rocchio Algorithm, Word Contribution

1 はじめに

近年，電子化されたテキストデータの増加により，ユーザにとって必要な情報を検索することが困難となり情報検索研究の社会的なニーズが高まっている．

代表的な情報検索モデルとして，ベクトル空間モデル (VSM)[1] がありVSMを用いた検索結果の改善手法として検索質問拡張手法がある．この手法は不十分な検索質問の情報を拡張させるというものでありRocchioの式を用いた適合性フィードバック手法が挙げられる [2][3][4]．しかしこの手法は文書単位でフィードバックを行い，初期クエリベクトルを更新している．そのため，同じ文書内にある単語には同じ修正値が与えられ，単語ごとの検索課題に対する影響は考慮できていない．この問題を解決した検索質問拡張手法に，個々の単語に単語寄与度という値を付与し初期クエリベクトルを更新する手法がある [5]．この手法の初期クエリベクトルを拡張させるための値は各単語の単語寄与度とある定数の積としている．ところが，検索精度に与える影響は単語ごとで異なるはずである．また，定数の最適値も明確ではない．

そこで本研究では，適合文書に特化した単語に適切な重みを与えるために単語の出現頻度を考慮した重みにより初期クエリベクトルを拡張する手法を提案する．また，この値は定数ではなく，出現頻度から得られた単語ごとの異なる値とする．さらに提案手法の有効性をベンチマークデータ (BMIR-J2)[6][7] を用いて検証する．

* 〒169–8555 東京都新宿区大久保 3–4–1 早稲田大学理工学部経営システム工学科 School of Science and Engineering, Waseda University 3–4–1 Okubo, Shinjuku-ku, Tokyo, 169–8555 Japan. E-mail: hiramatsu@hirasa.mgmt.waseda.ac.jp

2 従来の情報検索の研究

2.1 ベクトル空間モデル

まず検索に用いる単語の数を N とする．VSMでは，検索対象文書やユーザからの検索キーワードの集合である検索質問を N 個の単語のTF-IDF値を要素とする N 次元ベクトルで表現する．これらのベクトルをそれぞれ文書ベクトル，クエリベクトルという．VSMではユーザが与えたクエリベクトルに対して類似度の高い文書から検索結果として提示する．以下にVSMにおけるいくつかの定義を示す．

[定義 1:TF-IDF 値]
検索対象文書内の単語には特定性と網羅性を考慮したTF-IDF値と呼ばれる値 $w_{d_j}^{t_k}$ を与える．

$$w_{d_j}^{t_k} = \frac{f(t_k,d_j)}{F(d_j)} \times \left(1 + \log \frac{M}{df(t_k)}\right) \tag{1}$$

d_j : 検索対象文書 $(j = 1,2,\ldots,M)$
t_k : 検索対象文書集合に出現する単語 $(k = 1,2,\ldots,N)$
$f(t_k,d_j)$: 文書 d_j における単語 t_k の出現回数
$F(d_j)$: 文書 d_j の全単語数
$df(t_k)$: 単語 t_k が出現する文書数 □

[定義 2:文書ベクトル d_j]
文書 d_j の文書ベクトル d_j を次式で与える．

$$d_j = (w_{d_j}^{t_1}, w_{d_j}^{t_2}, \ldots, w_{d_j}^{t_N}) \tag{2}$$

□

[定義 3:クエリベクトル Q]
ユーザが最初に入力する検索キーワードは，q^{t_k} で表し，検索キーワードの集合からなる検索質問をベクトル表現したクエリベクトル Q は次式で表される．

$$Q = (q^{t_1}, q^{t_2}, \ldots, q^{t_N}) \tag{3}$$

$$q^{t_k} = \begin{cases} 0 & \text{単語 } t_k \text{ が検索キーワードでない} \\ 1 & \text{単語 } t_k \text{ が検索キーワードである} \end{cases}$$

□

[定義 4:クエリベクトル Q と文書ベクトル d_j の類似度 score(Q, d_j)]

検索質問 Q に対する文書 d_j の類似度をコサイン尺度として定義し，次式で与える．

$$\mathrm{score}(Q,d_j)=\frac{(Q,d_j)}{|Q||d_j|} \tag{4}$$

(Q,d_j) :クエリベクトル Q と文書ベクトル d_j の内積
$|Q|(|d_j|)$:ベクトル $Q(d_j)$ のノルム □

2.2 擬似フィードバック手法

擬似フィードバック手法は初期検索結果の文書に対し，システムが自動的に適合・不適合の判定を行い，これを用いてクエリベクトルを更新する手法である．ユーザ自身が適合・不適合の判定を行う Manual フィードバックという手法もあるが，この手法はユーザに負担にかかるため，本研究では擬似フィードバック手法のみを対象とする．代表的な擬似フィードバック手法として Rocchio の式によってクエリベクトルを更新する手法がある [2][3][4].

[Rocchio の式を用いた擬似フィードバック手法]

Rocchio の式を用いた擬似フィードバック手法では初期検索結果の類似度 $\theta(\theta\geq 0)$ 以上の文書を適合文書集合，類似度 θ 以下の文書を不適合文書集合と自動的に見なす．そして初期クエリベクトル Q を以下の式により更新する．

$$Q_{new}=Q+\lambda\frac{1}{M^+}\sum_{d^+\in R^+}d^+-\mu\frac{1}{M^-}\sum_{d^-\in R^-}d^- \tag{5}$$

Q_{new} : 更新後のクエリベクトル
$d^+(d^-)$: 適合 (不適合) 文書ベクトル
$\lambda(\mu)$: 適合 (不適合) の重要度を表すパラメータ
$R^+(R^-)$: 適合 (不適合) 文書集合 □

2.3 単語寄与度を用いた検索質問拡張手法

初期検索結果に対してシステムが自動的に判定した適合文書（類似度 θ 以上の文書）に出現する単語に単語寄与度を付与する．その値を用いて初期クエリベクルを拡張させ，新たなクエリベクトルを用いて再検索を行い，検索精度を向上させる手法が提案されている [5].

単語寄与度とは，ある単語を文書から除いた時に類似度がどのように変化するかを表す値と解釈できる．

[単語寄与度]

$$\mathrm{Cont}(t_l,Q,d^+)=\mathrm{score}(Q,d^+)-\mathrm{score}(Q'(t_l),d^{+'}(t_l)) \tag{6}$$

$Q'(t_l)$: 初期クエリベクトル Q の要素である，単語 t_l の値を 0 にしたベクトル
$d^{+'}(t_l)$: もとの適合文書ベクトル $d^{+'}$ の要素である，単語 t_l の値を 0 にしたベクトル
t_l : 適合文書に出現する単語 $(l=1,2,\ldots,N^+)$
N^+ : 適合文書に出現する単語数 □

式 (6) の単語寄与度を用いた検索の手順 [5] を以下に示す．

Step1 Q による初期検索

Step2 システムが自動的に適合の判定を行い，判定された適合文書中に出現する単語の単語寄与度を式 (6) を用いて求める．

Step3 Step2 で求めた単語寄与度を用いて式 (7) を計算する．

$$q_{new}^{t_l}=q^{t_l}+\log\left(\frac{M}{df(t_l)}\right)\times wgt\times\sum_{d^+}\mathrm{Cont}(t_l,Q,d^+) \tag{7}$$

$df(t_l)$: 単語 t_l が出現する文書数
wgt : 定数

Step4 式 (7) で得られた値うち適合文書にのみ出現する単語の値を初期クエリベクトルに加え、新たなクエリベクトル式 (8) を用いて再検索を行う．

$$Q_{new}=(q_{new}^{t_1},q_{new}^{t_2},\ldots,q_{new}^{t_l}\ldots,q_{new}^{t_N}) \tag{8}$$

□

単語寄与度の値が小さいため，単語寄与度同士の差は小さくなってしまう．差を拡大させるために定数 wgt をかけている．

$\mathrm{Cont}(t_l,Q,d^+)$ の絶対値が大きい時，つまりある単語を文書から除いた時に類似度が大きく変化する時，式 (7) は大きな値となる．そのような単語は有効な単語であると考えられる．

2.4 従来研究の問題点

2.4.1 VSM による初期検索の問題点

ユーザは不明確な事柄に対して検索要求が発生するため，得たい情報を検索質問として表現するのは困難である．そのため，初期クエリベクトルが不十分になり，ユーザが満足する検索結果は得られない [1]．そこで検索質問の拡張が必要であり Rocchio の式が提案された．

2.4.2 Rocchio の式を用いた擬似フィードバック手法の問題点

初期クエリベクトルに加えられる修正値は，式 (5) の第 2 項，第 3 項に見られるように，システムによって自動的に判定された適合文書集合，不適合文書集合の重心により求められている．つまり同じ文書内にある単語には同じ修正値が与えられ，各単語ごとの検索課題に対する影響は考慮できていないという問題が指摘されている [2][3][4].

2.4.3 単語寄与度を用いた検索質問拡張手法の問題点

2.3 節で述べた単語寄与度を用いた検索質問拡張手法では，式 (7) の wgt の値は定数であり，増加するにつれ

て $\mathrm{Cont}(t_l,Q,d^+)$ の初期クエリベクトルを更新する際の影響力が大きくなる．しかしその最適値は明確ではない．また，単語全てに同じ wgt がかけられているため，$\mathrm{Cont}(t_l,Q,d^+)$ は同じ影響力として初期クエリベクトルに加えられている．しかし単語によって $\mathrm{Cont}(t_l,Q,d^+)$ の検索課題に対する影響力は異なると考えられるので，定数 wgt を単語によって可変にすることで精度の向上が期待できる．

3 提案手法

提案手法では，従来で使われていた wgt という定数を改良し，適合文書に偏った単語には大きな重みを与えられるようにするため，出現頻度を用いた重み t_l-value を導入する．

適合文書に多く出現する単語はユーザの検索要求に関連のある単語である可能性が高い．しかし，このような単語はどのような文書にも出現する一般的な単語である可能性もある．そこで，適合文書に多く出現し，不適合文書にはあまり出現しない単語には大きな重みを与え，より適合文書に特化した単語を初期クエリベクトルから拡張できる手法を提案する．

[出現頻度を用いた重み]

$$t_l\text{-value} = \frac{Ucount(t_l)}{U} - \frac{Lcount(t_l)}{L} \tag{9}$$

$Ucount(t_l)$：適合文書内の単語 t_l の適合文書内での出現回数

$Lcount(t_l)$：適合文書内の単語 t_l の不適合文書内での出現回数

U：適合文書数

L：不適合文書数 □

式 (9) は適合文書に出現する単語の適合文書内での出現頻度と不適合文書内での出現頻度の差を表している．適合文書に多く出現し，不適合文書にあまり出現しない時，つまり適合文書に特化した単語には大きな値が与えられることになる．

[提案アルゴリズム]

Step1 Q による初期検索

Step2 式 (6) により，システムが自動的に判定した適合文書内の単語の単語寄与度を求める．

Step3 適合文書に特化した単語に適切な重みを与えるため，式 (9) を計算する．

Step4 Step2 で得られた単語寄与度と Step3 で得られた重みを用いて式 (10)，(11) を計算する．

(1)$\sum_{d^+}\mathrm{Cont}(t_l,Q,d^+) > 0$ の時

$$q_{new}^{t_l'} = q^{t_l} + \log\left(\frac{M}{df(t_l)}\right) \times \frac{t_l\text{-value} \times \sum_{d^+}\mathrm{Cont}(t_l,Q,d^+)}{\max(t_l\text{-value} \times \sum_{d^+}\mathrm{Cont}(t_l,Q,d^+))} \tag{10}$$

(2)$\sum_{d^+}\mathrm{Cont}(t_l,Q,d^+) < 0$ の時

$$q_{new}^{t_l'} = q^{t_l} + \log\left(\frac{M}{df(t_l)}\right) \times \frac{t_l\text{-value} \times \sum_{d^+}\mathrm{Cont}(t_l,Q,d^+)}{\min(t_l\text{-value} \times \sum_{d^+}\mathrm{Cont}(t_l,Q,d^+))} \tag{11}$$

Step5 式 (10)，(11) で得られた値のうち適合文書に偏って出現する t_l-value>0 の単語を初期クエリベクトルに加え，新たなクエリベクトル式 (12) で再検索を行う．

$$Q'_{new} = (q_{new}^{t_1'}, q_{new}^{t_2'}, \ldots, q_{new}^{t_l'} \ldots, q_{new}^{t_N'}) \tag{12}$$

□

単語寄与度の絶対値を求めた時に類似度の計算方法により，正と負の値で大きな差が生じる．しかし，正に大きな値をとる単語も，負に大きな値をとる単語も重要だと考えるため，式 (10)，(12) のように場合分けを行った．

また，$t_l\text{-value} \times \sum_{d^+}\mathrm{Cont}(t_l,Q,d^+)$ の値は非常に小さな値となる．そのため，初期クエリベクトルの要素の値とバランスをとるために $t_l\text{-value} \times \sum_{d^+}\mathrm{Cont}(t_l,Q,d^+)$ の値を正規化する．

定数 wgt を改良した重み t_l-value により適合文書に特化した単語の単語寄与度の影響力を大きくすることができ，検索精度の更なる向上が期待できる．

4 数値実験と考察

4.1 評価方法

提案手法の有効性を評価するために，数値実験を行った．評価方法としては，情報検索でよく用いられる以下の式 (13)，(14) で計算される再現率，適合率に対し，検索課題ごとの再現率 0.0,0.1,...,1.0 における適合率 (11 点適合率) とその加算平均である平均適合率で評価する．なお，平均適合率が大きいほど，ユーザの要求を満たす正解文書を上位に検索する能力があることを示す．

$$\text{再現率} = \frac{\text{検索された正解文書数}}{\text{正解文書数}} \tag{13}$$

$$\text{適合率} = \frac{\text{検索された正解文書数}}{\text{検索された総文書数}} \tag{14}$$

4.2 実験条件

評価データとして毎日新聞 1994 をもとにした BMIR-J2 テストコレクション [6] を使用し，その中から 10 課題に対して実験を行った．また，検索課題に対するユーザの要求に合った文書として，BMIR-J2 が提供する正解文書を使用した．

提案手法の比較対象として，VSM による初期検索 [1]，Rocchio の式を用いた擬似フィードバック手法 [2][3][4]，単語寄与度による検索質問拡張手法 [5] の 3 手法を使う．

パラメータとして初期クエリベクトルとの類似度 θ=0.3 以上の文書を擬似的に適合文書とした．また，単語寄与

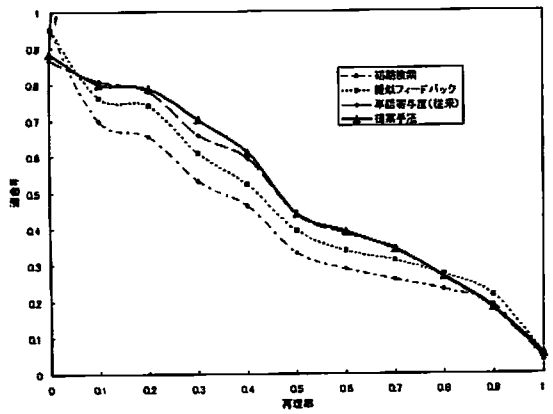

図 1: 全課題の 11 点平均適合率

表 1: 各手法の平均適合率

	初期検索	擬似フィードバック	単語寄与度	提案手法
平均適合率	0.4080	0.4606	0.4857	0.4953

表 2: 各手法の 30 位以内適合率

	初期検索	擬似フィードバック	単語寄与度	提案手法
平均適合率	0.6555	0.7098	0.7325	0.7504

表 3: 抽出された重要語上位 3 件（検索キーワード以外）

検索課題	手法	抽出された重要語上位 3 個		
農薬	提案手法	残留	輸入	マラチオン
	単語寄与度	残留	外来	検出
核兵器	提案手法	被爆	加害者	財産
	単語寄与度	被爆	加害者	財産
教育産業	提案手法	補助	公立	クーラー
	単語寄与度	クーラー	学習	補助
国連軍派遣	提案手法	空爆	セルビア	勢力
	単語寄与度	空爆	セルビア	安保
株価動向	提案手法	株式	平均	市場
	単語寄与度	平均	株式	市場
銀行	提案手法	阪神	仕方	主導
経営計画	単語寄与度	阪神	財務	八千島
映画	提案手法	館	スクリプター	労組
	単語寄与度	館	スクリプター	労組
女性の	提案手法	採用	企業	既婚
雇用問題	単語寄与度	採用	健全	毎年
日本企業	提案手法	台	販売	製品
の逆輸入	単語寄与度	台	製品	販売

度を用いた従来手法で用いられる wgt は $\mathrm{Cont}(t_l, Q, d^+)$ を正規化しているため 1 とした.

4.3 結果

図 1 に全課題 (10 課題) について得られた 11 点適合率の平均をグラフにした再現率・適合率曲線に示す. また表 1 に各手法の平均適合率, 表 2 に検索結果の上位 30 件までの平均適合率である 30 位以内適合率, 表 3 に, 従来の単語寄与度を用いた手法と提案手法により求められた単語のうち, 初期検索キーワード以外の新たにクエリベクトルに加わる値が大きい, 上位 3 件を示す.

4.4 考察

1. 図 1 より提案手法は, 従来手法 (擬似フィードバック手法, 単語寄与度を用いた手法) よりわずかではあるが高い適合率を得ることができたことがわかる.
2. 表 2 により 30 位以内適合率においても, 提案手法は従来手法を上回っていることが分かる. このことから、提案手法により検索結果の上位に, 従来手法より多くの正解文書を集めることが出来たことになる. 日常我々が行う検索では検索結果の上位に正解文書が数文書あれば十分であることを考慮すると, 30 位以内適合率で表 3 のような結果が得られたことで, 一般の検索の際に有効な手法であるといえる.
3. 従来手法も提案手法も各課題において, その分野に特化した単語が抽出できていることが分かる. 例えば, 検索課題「農薬」で抽出されたマラチオン (農薬の名前) は全部で 5 文書に出現し, その内ユーザの要求を満たす正解文書は 5 文書である. これは, マラチオンが正解文書の内容を的確に表す単語であることを表している. このように, 専門的な用語のような課題に対して十分な知識のないユーザには検索質問として入力することが困難な単語でも, 単語寄与度により適切な重みが与えられ, t_l-value により, さらに単語寄与度の効果を拡大させていると考えられる.

5 まとめと今後の課題

本研究では; 単語寄与度の考えをもとに VSM による初期検索で問題であった不十分な検索質問を補う検索質問拡張手法として, 出現頻度も考慮した単語寄与度を用いる手法を提案した. さらに数値実験により, 平均適合率, 上位文書に対して行った 30 位以内適合率の両方より従来手法より良い結果を得ることができた.

今後の課題として, 単語を拡張させる際に, 適合文書の内容を良く表す単語を抽出させるだけではなく, 適合文書の内容を表すには相応しくない不適切な単語をマイナスの要素として抽出させる手法を検討していきたい.

6 謝辞

著者の一人である平松は, 本研究を行うにあたり, 数多くのご助言, ご支援を賜りました早稲田大学平澤研究室の各氏に感謝いたします. 本研究の成果の一部は早稲田大学特定課題研究助成 No.2005B-189 による.

参考文献

[1] 北研二, "情報検索アルゴリズム," 共立出版, 2002 年
[2] Rocchio, J., "Relevance Feedback in Information Retrieval," *The SMART Retrieval System Experiments in Automatic Document Processing*, Prentice Hall Inc, 1971 年.
[3] 岸田和明, "文書検索におけるクエリーの拡張方法," 情報処理学会研究報告, No.67, Vol.2001, pp.55-62, 2001 年.
[4] Baeza-Yates, R., *Modern Information Retrieval*, Harlow, England, Addison-Wesley, Inc, 1999 年
[5] 帆足啓一郎, 松本一則, 井ノ上直己, 橋本和夫, "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, Vol.40, No.SIG8, pp.63-73, Nov.1999
[6] (社) 情報処理学会データベースシステム研究会, BMIR-J2, 新情報処理開発機構, 1998 年.
[7] 毎日新聞社, CD 毎日新聞'94, 日外アソシエーツ, 1995 年.