D-043

初期検索結果から抽出した単語を用いた擬似フィードック手法
Pseudofeedback technique that uses word extracted from preliminary search result

平松 丈嗣 † 松下 大輔 † 平澤 茂一 †
Jouji Hiramatsu Daisuke Matsushita Shigeichi Hirasawa

1. はじめに

近年の IT 技術の発展に伴い，個人で扱える電子化されたテキストデータが急激に増加している．その結果，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）がある．VSM は，文書と検索キーワードを単語の重みを要素とするベクトルで表現するモデルである．VSM を用いた検索結果の改善手法として適合性フィードバック手法がある．この手法はユーザが検索キーワードのいくつかの文書に対して適合 · 不適合の判定を行い，その判定結果をシステムに返すことにより検索精度を改善する手法である [1]．しかしこの手法はユーザに多大な負担をかけてしまう。

一方で，適合 · 不適合の判定を，ユーザでなくシステムが自動的に判定する (例えば，検索結果の上位文書を適合，下位文書を不適合と判定する) ことによって初期検索結果の精度を改善する擬似フィーバック手法の研究が行われている [2][3]．しかし，適合 · 不適合の判定が妥当であるかどうかは保証できない

本研究では，検索キーワードと関連した単語を自動的に抽出し，その単語に対して重要度に基づく適切な重みづけを与えることによってユーザに負担をかけることなく初期検索の精度を改善する手法を提案する．またテストデータ [6] に提案手法を適用し, 有効性を示す.

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[4] では，文書やユーザからの検索キーワードを N 個の単語の重みを要素とする 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．

[定義 1:単語の重み $w_{d_j}^{t_k}$]
索引語の重み $w_{d_j}^{t_k}$ は TF · IDF 値で与えられる．

$$w_{d_j}^{t_k} = (f(t_k, d_j)/F(d_j)) \times (1 + \log(M/df(t_k))) \quad (1)$$

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 $\boldsymbol{d_j}$ を次式で与える．

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

[定義 3:クエリベクトル Q] □
検索質問は検索キーワードの集合として与えられ, 次式のクエリベクトルで表される．

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

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

[定義 4:クエリベクトル Q と文書 d_j の類似度 $score(\boldsymbol{Q}, \boldsymbol{d_j})$]
検索質問 $\boldsymbol{Q}$ に対する文書 d_j のスコアを次式で与える．

$$score(\boldsymbol{Q}, \boldsymbol{d_j}) = (\boldsymbol{Q}, \boldsymbol{d_j})/|\boldsymbol{Q}||\boldsymbol{d_j}| \quad (4)$$

$(\boldsymbol{Q}, \boldsymbol{d_j})$: クエリベクトル $\boldsymbol{Q}$ と文書ベクトル $\boldsymbol{d_j}$ の内積
$|\boldsymbol{Q}|$: ベクトル $\boldsymbol{Q}$ 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．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．

[Rocchio の式を用いた自動フィードバック手法]
Rocchio の式を用いたフィードバック手法 [2][3] では初期検索結果の上位 M^+ 文書を適合文書，下位 M^- 文書を不適合文書と見なす．ここで元のクエリベクトル $\boldsymbol{Q}$ を以下の式により更新する．

$$\boldsymbol{Q_{new}} = \boldsymbol{Q} + \lambda \frac{1}{M^+} \sum_{\boldsymbol{d}^+ \in R^+} \boldsymbol{d}^+ - \mu \frac{1}{M^-} \times \sum_{\boldsymbol{d}^- \in R^-} \boldsymbol{d}^- \quad (5)$$

$\boldsymbol{Q_{new}}$: 更新後のクエリベクトル
$\boldsymbol{d}^+(\boldsymbol{d}^-)$: 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) 文書集合である □

2.3 従来手法の問題点

ユーザは不明確な事柄に対して文書検索を行うため，得たい情報を検索キーワードとして表現するのは困難である．また擬似フィードバック手法は単語単位ではなく，文書という大きな規模でフィードバックしているため，上位文書（下位文書）に含まれる不正解文書（正解文書）の誤判定の影響を文書単位で大きく受けてしまう．

3. 提案手法

本研究では，ユーザが与えた検索質問と関連のある単語を見つけ，それに，適切な修正値を与えることにより，クエリベクトルの更新を行う手法を提案する．

3.1 検索キーワードと関連のある単語の抽出

初期検索結果は，必ずしもユーザが求めている文書集合ではなく，ユーザが与えたクエリベクトルと類似度が高い文書集合である．そこでクエリベクトルと類似度の高い文書に出現する単語を，検索キーワードと関連のある単語とみなす．そしてこの単語に修正値を与える．ここではクエリベクトルとの類似度に対し閾値 $\theta(>0)$ を用いて，検索キーワードと関連のある単語の抽出を行う．

3.2 単語の修正値の計算

更新前のクエリベクトルと類似度が高い文書に出現する単語は，検索キーワードと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文書にも出現するような単語である可能性もある．そこで上位文書に出現する単語 t_k の出現率と上位文書以外の文書に出現

†早稲田大学大学院理工学研究科経営システム工学専攻

する単語 t_k の出現率の差分をとる．また検索課題によってクエリベクトルとの類似度が異なる。そこで上位文書に出現した単語の出現率と上位文書の類似度の平均の積をとり正規化を行う. その値から上位文書以外の文書に出現する単語 t_k の出現率の差分をその単語の修正値とする．単語 t_k の修正値 r^{t_k} は式 (6) のように定義する．

[定義 5:単語 t_k の修正値 r^{t_k}]

$$r^{t_k} = \frac{\sum_{j=1}^{U} sortscore(\boldsymbol{Q}, \boldsymbol{d_j})}{U} \cdot \frac{Ucount\ \ (t_k)}{U} - \frac{Lcount\ \ (t_k)}{T-U} \quad (6)$$

r^{t_k} : 単語 t_k の修正値
$sortscore(\boldsymbol{Q}, \boldsymbol{d_j})$:類似度 θ 以上の文書の類似度
$Ucount\ \ (t_k)$:$\boldsymbol{Q}$ との類似度 θ 以上の文書に出てくる単語 t_k の出現回数
$Lcount\ \ (t_k)$:$\boldsymbol{Q}$ との類似度 θ 以下の文書に出てくる単語 t_k の出現回数
U : $\boldsymbol{Q}$ との類似度 θ 以上の文書数
T : 全文書数 □

この r^{t_k} の値が大きい単語ほど，検索キーワードと関連があり，検索する上で重要な単語であると考えられる．また類似度の平均と上位文書の出現率の積をとることで，クエリベクトルとの類似度が高い文書に出現する単語には高い値が与えられる．

3.3 クエリベクトルの更新

式 (7) によりクエリベクトルを更新し検索を行う．

$$\boldsymbol{Q'} = (q^{t_1} + r^{t_1}, q^{t_2} + r^{t_2}, \ldots, q^{t_N} + r^{t_N}) \quad (7)$$

4. 評価実験

提案手法の有効性を示すために評価実験を行う．実験データとして毎日新聞 1994[5] をもとにした BMIR-J2 テストコレクション（5,080 文書）[6] を用いた．また使用する検索課題の数は 10 課題とした．

擬似フィードバック手法において，θ≥0.3 の文書を適合と判定し，θ<0.3 を不適合と判定した．また，提案手法の閾値は θ=0.3 とした．

評価方法は，検索課題ごとの平均適合率と全課題の 11 点適合率の平均を用いる．

5. 結果と考察

5.1 結果

表 1 に各課題の初期検索，従来手法，提案手法の平均適合率. 図 1 に全課題の 11 点適合率の平均. 表 2 に抽出された単語とその修正値の上位 4 個を示す.

表 1: 各課題の平均適合率

検索課題	初期検索	従来手法	提案手法
農薬	0.381	0.561	0.578
飲料品	0.305	0.306	0.305
核兵器	0.604	0.661	0.677
教育産業	0.295	0.309	0.356
国連軍派遣	0.353	0.450	0.531
株価動向	0.583	0.634	0.669
銀行の経営計画	0.272	0.312	0.368
映画	0.527	0.539	0.544
女性の雇用問題	0.411	0.459	0.477
日本企業による逆輸入	0.351	0.378	0.468
平均	0.405	0.461	0.497

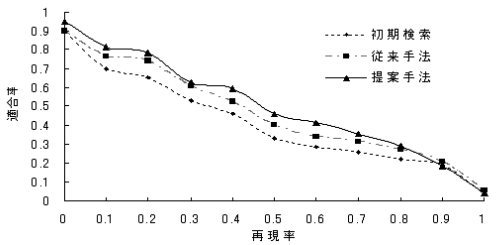

図 1: 全課題の 11 点平均適合率

表 2: 抽出された単語とその修正値

検索課題	抽出された単語とその修正値（上位 4 個）			
農薬	マラチオン	子孫	殺虫剤	残留
	0.3704	0.3698	0.3698	0.3668
核兵器	投稿	改訂	加害者	本欄
	0.3240	0.3240	0.3225	0.3221
教育産業	**学校**	公立	生徒	高校
	0.3772	0.2321	0.2314	0.2253
国連軍派遣	**国連**	セルビア	勢力	攻撃
	0.2811	0.2398	0.2297	0.2005
株価動向	**株価**	株式	証券	終値
	0.3613	0.2670	0.2564	0.2514
銀行経営計画	不良	債権	**経営**	**銀行**
	0.3291	0.3279	0.2885	0.2781
映画	**映画**	映画館	入場	監督
	0.5369	0.3265	0.3230	0.3163
女性の雇用問題	**雇用**	労働	**女性**	**就職**
	0.1723	0.1706	0.1696	0.1533
日本企業による逆輸入	**輸入**	**逆輸入**	最高	円高
	0.2539	0.2319	0.1828	0.1638

5.2 考察

1. 表 1 より 10 課題中 9 課題で提案手法は従来手法を上回っていることが分かる。
2. 図 1 より再現率が 0.8 以下の時，提案手法は従来手法の性能を上回っている. 通常の検索ではユーザは検索結果の上位にのみ興味があることより，本研究の提案手法は有効な手法であるといえる.
3. 表 2 より検索課題に関連した有効な単語を抽出できていることが分かる. 例えば,「農薬」に関する文書が欲しい時,「マラチオン」（農薬の名前）という単語の抽出に成功している. この単語は専門的な言葉であり, これから調べようとしているユーザにはおそらく思いつかないキーワードであろう. よってこの手法はユーザの検索課題に対する知識が少ない時ほど有効な手法であると考えられる.
4. 検索課題の「映画」は従来手法も提案手法も初期検索から検索精度の向上があまり見られなかった. この原因として, 新聞データにある「映画」関する文書はユーザにとって一般的で, 初期検索の段階でユーザの持っている情報が明確であり, 文書の特徴を現す検索質問を得ることができたためと考えられる.
5. 提案手法が従来手法を上回ったのは提案手法は単語ごとにフィードバックすることで，従来手法で起こりえる文書の誤判定の影響を小さくしたためと考えられる.

6. むすび

本研究は擬似フィードバックの考えをもとに，ユーザにとって関連のある重要な単語を自動抽出し，その単語を用いたクエリベクトルで検索することにより初期検索精度を向上させることができた.

本研究で行われた手法は特徴的な単語が使用される課題に対しては有効な結果得られた. このことから論文検索のような，普段我々が情報が少ない状態で検索するであろう分野への応用を考えていきたい.

参考文献

[1] Rocchio, J.,Relevance Feedback in Information Retrieval, *The SMART Retrieval System Experiments in Automatic Document Processing*, Prentice Hall Inc, 1971 年.

[2] Baeza-Yates, R., *Modern Information Retrieval*, Harlow,England,Addison-Wesley,Inc, 1999 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] 北研二，情報検索アルゴリズム，共立出版, 2002 年

[5] 毎日新聞社, CD 毎日新聞’94, 日外アソシエーツ, 1995 年.

[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.