ユーザにとって潜在的に重要な単語を用いた対話的文書検索

Interactive Document Retrieval Using the Important Word Potentially for a User

松下 大輔 † **足立 鉱史** † **平澤 茂一** †

Daisuke Matsushita Hiroshi Adachi Shigeichi Hirasawa

1 はじめに

近年のインターネットに代表されるように，個人で扱える電子化された文書データが増加している．その結果，ユーザにとって必要な情報を効率的に取り出せる情報検索システムに対するニーズがますます高まってきている．

一般的な情報検索システムとして，対象とする文書データを索引語の多次元ベクトルで表現するベクトル空間モデル（VSM）が広く利用されている．VSM を用いた検索結果に対して，ユーザが適合 · 不適合の情報をシステムにフィードバック（適合フィードバック）することで対話的に検索精度を向上させている [1].

本研究では，ある文書においてユーザにとって重要な索引語はユーザが入力した検索キーワードの前後に現れると考え，これを考慮した検索手法を提案する．そして，これを実装し，ベンチマークデータ（BMIR-J2）を用いて有効性を示す.

2 従来の情報検索手法

2.1 ベクトル空間モデル

VSM における検索では，文書やユーザからの検索質問は検索対象文書集合の索引語数を次元とした文書ベクトルとクエリベクトルで表現する [2].

[定義 1:索引語の重み $w_{d_j}^{t_k}$]

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索引語 $(k = 1, 2, \ldots, N)$

$f(t_k, d_j)$: 文書 d_j における索引語 t_k の出現回数

$F(d_j)$: 文書 d_j の全索引語数

$df(t_k)$: 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．

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

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

□

[定義 4:文書 d_j のスコア $score(\boldsymbol{Q}, \boldsymbol{d_j})$]

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．

$$score(\boldsymbol{Q}, \boldsymbol{d_j}) = \sum_{k=1}^{N} (q^{t_k} \times w_{d_j}^{t_k}) \quad (4)$$

□

2.2 適合フィードバック

適合フィードバックは，初期検索結果で得た文書集合の一部に対し，ユーザが適合 · 不適合の判定を行うことで検索システムの精度を対話的に改善する手法である．以下に，適合フィードバックのシステムの概要を説明する．

1) ユーザが検索キーワードを入力する．
2) システムは，各文書ごとにクエリベクトルに対するスコアを計算し，その降順に並べた文書リスト（初期ランキング）を出力する．
3) ユーザは上位数文書（20 文書程度）に対し，適合 · 不適合のラベル付けの判定を行い，システムにフィードバックを行う．
4) システムは，フィードバック情報をもとにクエリベクトルを更新する．
5) システムは，4) で得た新しいクエリベクトルを用いて，各文書のスコアを再計算し，新しいランキングをユーザに提示する． □

2.3 Rocchio アルゴリズム

古くから用いられるクエリベクトルの更新方法として Rocchio アルゴリズム [3] が挙げられる．新しいクエリベクトル $\boldsymbol{Q}_{new}$ は，適合文書集合を表すベクトル群の重心と，不適合文書を表すベクトル群の重心との差分ベクトルとする．

[定義 5:Rocchio のクエリベクトル更新式]

$$\begin{aligned} \boldsymbol{Q}_{new} &= (q_{new}^{t_1}, q_{new}^{t_2}, \ldots, q_{new}^{t_N}) \\ &= \frac{1}{|D^+|} \sum_{d_j \in D^+} \boldsymbol{d}_j - \frac{1}{|D^-|} \sum_{d_j \in D^-} \boldsymbol{d}_j \end{aligned} \quad (5)$$

$D^+(D^-)$: ユーザが (不) 適合と判断した文書集合

$|D|$: 集合 D の要素数 □

2.4 従来手法の問題点

VSM に基づいた検索システムは，表現能力上，次のような問題点が存在する．

- VSM では特定の単語間の共起関係に着目していない
- 適合文書における特徴的な単語やその組み合わせが表現できていない

†早稲田大学大学院理工学研究科経営システム工学専攻

3 提案手法

ユーザが入力した検索キーワードの前後に出現する索引語は重要であると考えられる．そこで，本研究ではユーザが入力した検索キーワードをもとに，検索キーワードと適合文書中の索引語を組み合わせて新たな適合文書の検索に有効な共起関係を発見し，この共起関係を利用した検索手法を提案する．

3.1 重要語抽出のための定義式

[定義 6:索引語 t_k の重要度 $tscore(t_k)$]

$$tscore(t_k) = \max_{d_j \in D^+} (q_{new}^{t_k} \times w_{d_j}^{t_k}) \quad (6)$$

□

式 (6) で新しいクエリベクトル $\boldsymbol{Q}_{new}$，文書ベクトル $\boldsymbol{d}_j$ の両方において重要な索引語を抽出し，その索引語はユーザにとって重要となる単語と考えられる．

3.2 近接単語対の作成方法

まず，重要度 $tscore(t_k)$ 上位 150 個の索引語を抽出し，これを重要語集合 W とする．なお，重要語抽出数 150 個は予備実験により経験的に与えられたものである．

次に近接単語対集合 R を作成する．近接単語対とはある適合文書中の同一文内で共起するユーザが入力した検索キーワードと重要語の組である．ただし，作成された近接単語対が不適合文書中に存在するのは好ましくないため，その場合はその近接単語対を削除する．ここで，同一文内という制約のない場合は，検索キーワードと重要語の組の数が多くなるため，適合文書を特徴付けることができないと考えられる．

3.3 近接単語対を利用した検索

作成された近接単語対が 1 つでも出現する文書を抽出し，これを優遇文書集合とする．次に，新しいクエリベクトル $\boldsymbol{Q}_{new}$ を用いて，優遇文書集合に対して再検索を行いランキングを出力する．

4 シミュレーションと考察

4.1 評価方法

提案手法の効果を検証するため，シミュレーションによる実験を行った．評価方法としては，以下の式 (7),(8) 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 $0.0, 0.1, \cdots, 1.0$ における適合率（11 点適合率）とその平均値（平均適合率），また近接単語対の効果がより顕著に現れる検索結果上位での適合率を見るため 20 位以内適合率を用いる．

$$\text{再現率} = \frac{\text{検索された正解文書数}}{\text{正解文書数}} \quad (7)$$

$$\text{適合率} = \frac{\text{検索された正解文書数}}{\text{検索された総文書数}} \quad (8)$$

4.2 シミュレーション条件

評価データとして毎日新聞 1994[4] をもとにした BMIR-J2 テストコレクション（5,080 文書）[5] を用いた．また評価対象 10 課題に対するユーザの適合・不適合の判定には BMIR-J2 が提供する正解文書，不正解文書を使用した．実験では，20 文書をフィードバックしてクエリベクトルの更新を行った．

4.3 結果

11 点適合率の全課題について平均をとり，これをグラフにした再現率・適合率曲線を図 1 に示す．また，平均適合率と 20 位以内適合率の全課題についての平均を表 1 に示す．

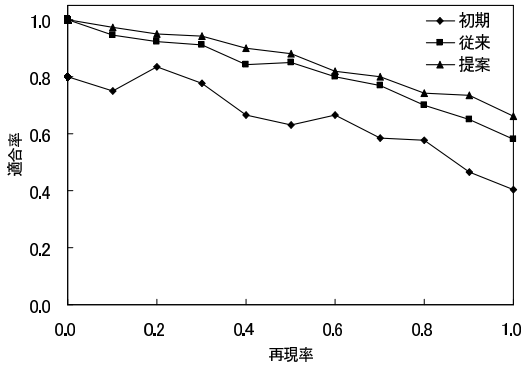

図 1: 再現率・適合率曲線

表 1: 各手法の平均適合率と 20 位以内適合率

	初期	従来	提案
平均適合率	0.650	0.801	0.832
20 位以内適合率	0.498	0.742	0.786

4.4 考察

図 1 と表 1 より提案手法は適合率において，従来手法より優れた結果を得ている．特に，再現率が 0.3 以上から近接単語対の効果が徐々に出てきていることが分かる．これは，提案手法において再現率が 0.3 を超えるあたりから，ユーザにとって潜在的に重要な単語を含む文書が，従来手法に比べて検索結果の上位に移動したためと考えられる．同様のことが，再現率が低い状態においても言える．逆に，再現率が 0.3 以下の状態では，従来手法と提案手法は適合率において，大差ない結果となった．これは，従来手法においても検索結果上位では，近接単語対を含む文書が多かったためと考えられる．

5 まとめと今後の課題

VSM において表現が不十分な単語間の共起関係を考慮し，提案手法において従来手法より高い適合率を得ることができた．

今回は近接単語対を索引語単位で作成したが，今後は，重要語抽出について日本語の係り受け関係を利用して重要語を抽出する手法などを検討していきたい．

参考文献

[1] 岡部正幸,“ 関係学習を用いた対話的文書検索 ”, 人工知能学会論文誌,No.6,Vol.16,pp.139-146,2001 年.
[2] 徳永健伸, 情報検索と言語処理, 財団法人東京大学出版会,1999 年.
[3] Rocchio,J.,Relevance Feedback in Information Retrieval,*The SMART Retrieval System Experiments in Automatic Document Processing*,Prentice Hall Inc,1971 年.
[4] 毎日新聞社,CD 毎日新聞'94, 日外アソシエーツ,1995 年.
[5] (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.