E-001

文脈関連度による検索質問拡張手法の改良
Query Expansion using Mutual Contextual Document Relevance

坂口 朋章 † Tomoaki Sakaguchi　**平松 丈嗣** † Jouji Hiramatsu　**平澤 茂一** † Shigeichi Hirasawa

1. はじめに

インターネットの普及に伴い，電子データの量が急増しており，必要な情報の抽出を困難としている．そのため近年，効率的な情報検索技術が注目を集めている．

代表的な情報検索のモデルに，ベクトル空間モデル（Vector Space Model：VSM）がある．VSM は，文書と検索質問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結果のいくつかの文書に対して適合・不適合の判定を行い，検索精度を改善する手法である [1]．一方，適合・不適合の判定を，システムが自動的に判定することで，初期検索結果の精度を改善する自動フィーバック手法の研究も行われている [2].

本研究では，与えられた検索質問に対して拡張した文脈関連度を用いて，関連性のある単語の集合を文書集合の中から抽出し，自動的に検索質問を拡張する手法を提案する．そして，提案手法をベンチマークデータ [3] に適用し，実験により有効性を示す.

2. 従来の情報検索の研究

2.1 ベクトル空間モデル (VSM)

検索に用いる文書の数を M，単語の数を N とする．VSM では，文書やユーザからの検索質問を 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する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 $w_{d_j}^{t_k}$，文書ベクトル $\boldsymbol{d}_j$]

単語 t_k の文書 d_j における重み $w_{d_j}^{t_k}$ は (1) 式の TF・IDF 値で与えられる．また，文書 d_j の文書ベクトル $\boldsymbol{d_j}$ は (2) 式で与えられる．

$$w_{d_j}^{t_k} = (f_{d_j}^{t_k}/F(d_j)) \times (1 + \log(M/df(t_k))) \tag{1}$$

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

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 ($k = 1, 2, \ldots, N$)
$f_{d_j}^{t_k}$：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: クエリベクトル Q]

検索質問は次式のクエリベクトル $\boldsymbol{Q}$ で表される．

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

$$q^{t_k} = \begin{cases} 0, & \text{単語 } t_k \text{ が検索語でない} \\ 1, & \text{単語 } t_k \text{ が検索語である} \end{cases}$$

[定義 3: 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 の類似度を次式で与える．

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

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

†早稲田大学理工学部経営システム工学科

2.2 自動フィードバック手法

自動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．自動フィードバック手法として Rocchio の式により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][2]

Rocchio の式を用いたフィードバック手法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し，$\boldsymbol{Q_{new}}$とする．

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

$\boldsymbol{d}^+(\boldsymbol{d}^-)$：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

2.3 関連語を用いた検索質問拡張手法

自動フィードバック手法では文書単位で適合・不適合を判断している．それに対し，各単語が初期検索語に対して関連語かどうかを評価し，評価値が上位の単語（関連語）を初期クエリベクトルに加える手法があり，関連語を用いた検索質問拡張手法と呼ばれる．関連性の評価値のひとつに文脈関連度がある．

[文脈関連度][4]

文脈関連度は，単語の重みと類似度を用いて検索質問との関連性を与える評価値である．類似度が上位の文書に出現している単語は，検索質問との関連性が高いとみなし，大きい重みを付与する．文脈関連度 $ncdr(\boldsymbol{Q}, t_k)$ は次式で定義される．

$$ncdr(\boldsymbol{Q}, t_k) = \frac{\sum_{j=1}^{M} w_{d_j}^{t_k} score(\boldsymbol{Q}, \boldsymbol{d_j})}{\sum_{j=1}^{M} w_{d_j}^{t_k}} \tag{6}$$

$ncdr(\boldsymbol{Q}, t_k)$：文脈関連度

この文脈関連度が高いほど検索質問に関連のある単語となる．

3. 提案手法

3.1 従来の手法の問題点

1. 従来の自動フィードバック手法では文書の適合・不適合判定をシステムが行う手法で，ユーザが得たい情報を検索質問として表現する手法ではない.
2. 文脈関連度は各文書内での単語の重みを用いるが，初期検索語の各文書内での重みは利用しない．そのため，初期検索語との関連性を表現するのに不十分である.

本研究では，初期検索質問と関連のある単語を見つけ，これらの単語をもとにクエリベクトルを更新する手法を提案する．この手法は，従来用いられていた文脈関連度を拡張したもので，検索結果の精度向上が目的である.

3.2 検索語と関連のある単語の評価値

初期検索語の各文書内での重みを利用することで，初期検索語と単語の関連性を正確に評価できると考えられる．そこで，従来の文脈関連度に初期検索語の重みを評価した値を加えた相互文脈関連度を提案する．

[**提案する相互文脈関連度**]

$$cncdr(\boldsymbol{Q}, t_k) = ncdr(\boldsymbol{Q}, t_k) + \alpha \sum_{i=1}^{N} \frac{q^{t_i} \times \sum_{j=1}^{M} w_{d_j}^{t_k} score(\boldsymbol{Q'_i}, \boldsymbol{d_j})}{\sum_{j=1}^{M} w_{d_j}^{t_k}} \quad (7)$$

$cncdr(\boldsymbol{Q}, t_k)$：単語 t_k の相互文脈関連度

$\boldsymbol{Q'_i} = (\underbrace{0, 0, \ldots, 0}_{(i-1)\text{ 個}}, 1, 0, \ldots, 0)$

α = 提案手法で追加した第 2 項の重み ($\alpha \geq 0$)

3.3 再検索

再検索の際には，相互文脈関連度が上位の単語をクエリベクトルに追加することにより，検索質問を拡張する．

[**検索追加語 t_k に加える重み $w_{add}^{t_k}$**]

$$w_{add}^{t_k} = \frac{cncdr(\boldsymbol{Q}, t_k)}{\max_i cncdr(\boldsymbol{Q}, t_i)} \quad (8)$$

$\max_i cncdr(\boldsymbol{Q}, t_i)$：$cncdr(\boldsymbol{Q}, t_i)$ の最大値

検索質問を拡張した後，更新したクエリベクトルで再検索を行い，ユーザに結果を提示する．

4. 評価実験

4.1 実験方法

提案手法の有効性を示すために評価実験を行う．(1) 初期検索結果（初期），(2)Rocchio の式を用いた自動フィードバック手法（AUTO），(3) 文脈関連度による検索質問拡張手法（従来），(4) 提案手法（提案）で実験を行い，結果を比較する．(　) は図表中の略称を表す．実験データとして毎日新聞 1994 [5] をもとにした BMIR-J 2 テストコレクション（5,080 文書）[3] を用いた．また使用する検索課題の数は 10 課題とした．

実験に用いるパラメータは，予備実験により以下のように定めた．自動フィードバック手法で適合とする文書数は 1 位〜20 位の 20 件，不適合とする文書数は 51 位〜100 位の 50 件とした．文脈関連度と相互文脈関連度を用いた検索での追加検索語の候補は初期検索結果の上位 30 件に出現した単語とした．

4.2 評価方法

評価方法として，以下の式 (9)，(10) 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) とその平均値 (平均適合率) を用いる．

$$\text{再現率} = \frac{\text{検索された正解文書数}}{\text{正解文書数}} \quad (9)$$

$$\text{適合率} = \frac{\text{検索された正解文書数}}{\text{検索された総文書数}} \quad (10)$$

5. 結果と考察

5.1 結果

パラメータは予備実験により決定した．$\alpha = 7$，300 語追加した際の 11 点適合率の全課題についての平均値を図 1 に，全課題についての 11 点平均適合率の平均値を表 1 に示す．

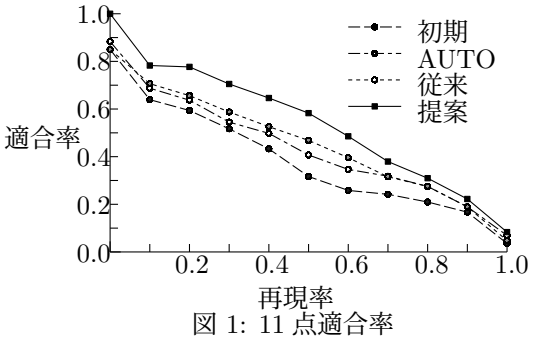

図 1: 11 点適合率

表 1: 平均適合率の比較

	初期	AUTO	従来	提案
平均適合率	0.378	0.437	0.455	0.541

5.2 考察

1. 表 1 より 10 課題の検索結果の平均した平均適合率は従来手法を上回っている．
2. 再現率の低い場合，すなわち検索結果が上位の文書のみを考慮した場合に対して従来よりも良い性能を示している．これは，初期検索語と関連の高い単語を抽出できたためだと思われる．例えば,「飲料品」の検索課題では従来手法では上位にある「ワイン」・「清酒」などの単語が提案手法により上位になり，従来手法では下位にある「農家」・「化粧水」などの単語が上位になった．
3. 検索追加語は 200 語前後，$\alpha = 7$ 付近で最も適合率が高くなった．検索追加語の語数や α の値を変化させた場合でも，従来手法よりも優れた結果を示した．このようなパラメータはデータセットに依存していると考えられるので，実問題に適用する際には予備実験等を行う必要がある．
4. 研究室に保存されている文書の検索に本手法を適用したところ，従来手法と提案手法にあまり差が見られなかった．これは，研究室内の文書の話題が限定されており，追加検索語による話題の特定の影響が小さかったからである．
5. 従来手法に比べて計算量は数倍になっている．

6. むすび

本研究では検索質問拡張手法に対して相互文脈関連度という評価式を提案し，実験により有効性を示した．

また，実問題にも適用し，有効性を確認する必要がある．

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] 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.

[3] (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.

[4] 佐々木 稔, 北 研二, “文脈関連度による検索質問の関連語抽出”, 言語処理学会第 7 回年次大会, pp.105-108, 2001 年.

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