データベースシステム 119−51
（１９９９．７．２２）

関連性の重ね合わせモデルに基づく問い合わせ表現の拡張

金沢 輝一† 高須 淳宏‡ 安達 淳‡

†東京大学大学院工学系研究科 ‡学術情報センター研究開発部

〒 112-8640 東京都文京区大塚 3-29-1 学術情報センター

TEL: 03–3942–6995 E-mail: {tkana, takasu, adachi}@rd.nacsis.ac.jp

筆者らは情報検索における自然言語の意味曖昧性への対処として関連性の重ね合わせモデル (RS モデル) を提案している．この手法は，索引語の重要度計算を tf·idf などの手法より高い精度で行うことができる．一方，検索者が検索対象に不慣れで曖昧な問い合わせ表現を入力した場合には関連語を補う query expansion が有効だが，既存の自動化された手法では元の表現による検索結果がある程度の精度を有していない場合には効果が得られないという問題点も指摘できる．本稿では query expansion の自動化と RS モデルを融合させ，文書関連性に基づくクラスタから重要語を選択することで，不慣れな検索者の問い合わせからも高精度に検索を行う手法を提案し，その有効性を示す．

キーワード 情報検索，ベクトル空間モデル，文章ベクトル拡張，
query expansion，RS モデル，NTCIR

Query Expansion with the Relevance-based Superimposition Model

Teruhito KANAZAWA† Atsuhiro TAKASU‡ Jun ADACHI‡

† Graduate School of Engineering, University of Tokyo

‡ R & D Department, NACSIS (National Center for Science Information Systems)

NACSIS, 3-29-1, Otsuka, Bunkyo-ku, Tokyo 112-8640, JAPAN

TEL: +81–3–3942–6995 E-mail: {tkana, takasu, adachi}@rd.nacsis.ac.jp

We have proposed a Relevance-based Superimposition (RS) model to solve the problems of semantic ambiguity on information retrieval. This method enables more accurate estimation of the weights of index terms than conventional methods such as tf·idf. Query expansion, which adds searching terms to the query, is considered useful when a novice user inputs an ambiguous query. However, most of automatic query expansion methods cannot achieve sufficent effectiveness when the quality of search results by the original query is very low. In this paper, we propose an automatic query expansion method which can be combined with the RS model and choose the relative searching terms from document clusters constructed based on the relevance of documents. We show the effectiveness of the proposed method through experiments.

Keyword information retrieval, vector space model, document vector expansion,
query expansion, RS model, NTCIR

1 はじめに

情報検索の精度を下げる要因として，問い合わせ表現 (query) と検索対象の文書が共に意味曖昧性を持っており，不慣れな検索者が対象に適した問い合わせを入力できないことが挙げられる．

筆者らはベクトル空間モデル上で文書側のベクトルを関連性に基づいて拡張することで検索精度を向上する手法を提案してきた [1, 2]．一方，関連語を追加することで query ベクトルを修正する query expansion が従来から行われてきているが，自動化された手法は検索者の意図を十分反映できず，人手を介する手法に比べて性能向上への寄与が小さいとされている [3]．

本論文では relevance feedback による関連語抽出が予備検索の上位候補から適切な語を選択できない場合に着目し，文書関連性によって形成したクラスタを用いて関連語を選択する手法を提案する．

2 章では，我々が提案している RS モデルの定義を述べ，評価用検索システムの実装について説明する．次に 3 章で今回の実験に採用した query expansion 手法の特性を調べ，その問題点を指摘する．4 章では前章で述べた問題点を克服すべく，RS モデルと組み合わせた自動 query expansion 手法を提案し，その効果を評価，考察する．

2 RS モデル

関連性の重ね合わせモデル (RS モデル) は，筆者らが提案している意味曖昧性への対策手法である．これは，検索対象の文書間に存在する関連性に基づき非排他的な文書集合を作り，これを解析することで文書ベクトルを拡張するというものである．以下にその定義を述べる．

2.1 非排他型クラスタの生成

文書群 $\{\mathrm{d}_1, \mathrm{d}_1, \cdots, \mathrm{d}_n\}$ で構成されたデータベースを仮定する．また，各々の文書に対応する文書ベクトルを $\{\boldsymbol{d}_1, \boldsymbol{d}_1, \cdots, \boldsymbol{d}_n\}$ と定義する．RS モデルでは文書を非排他型クラスタ $\{C_1, C_2, \cdots, C_m\}$ に分類する．図 1 は分類の例で，クラスタはキーワードによって形成されている．図中のデータベースには A と B の 2 つのキーワードが存在し，キーワード A が付与された文書で構成されたクラスタ C_A と，キーワード B が付与された文書で構成されたクラスタ C_B が形成されている．図中，文書 d_1 はキーワード A のみが付与されているのでクラスタ C_A に属する．一方文書 d_3 はキーワード A, B ともに付与されているので C_A と C_B の両方に属する．

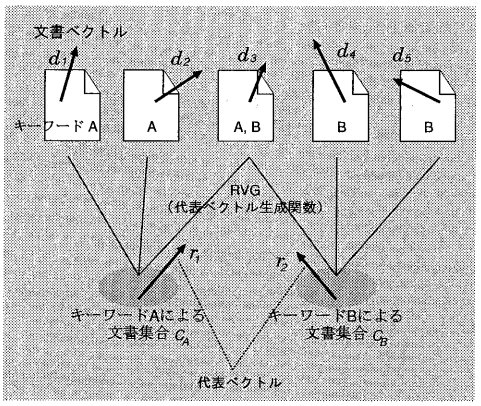

図 1　キーワードによる関連文書クラスタの生成

2.2 代表ベクトルの生成

RS モデルによる文書ベクトルの拡張は，クラスタの代表ベクトル生成と，代表ベクトルを用いての文書ベクトルの実質的な拡張の 2 つの段階を経て行われる．

まず最初の段階として，文書クラスタごとに代表となる特徴ベクトルを生成する．このベクトルは文書ベクトルと同じ特徴空間内のベクトルであり，同数の次元を持つ．クラスタ C の代表ベクトル $\boldsymbol{r}$ は C に属する全文書のベクトルを引数とする代表ベクトル生成関数によって生成される．これまでの実験 [2] によると，最も良い性能を示す代表ベクトル生成関数は，Root-Mean-Square を用いたもので，代表ベクトル $\boldsymbol{r}$ の第 i 要素を次のように求める関数である．

$$\sqrt{(\frac{1}{|C|}\sum_{\mathrm{d}_j \in C} d_{j,i})^2} \tag{1}$$

ただし，$d_{j,i}$ は文書 d_j の文書ベクトル $\boldsymbol{d}_j$ の第 i 要素である．

2.3 文書ベクトルの修正

次に，代表ベクトルを用いて各文書のベクトルを拡張する．文章が属する全ての文書群の代表ベクト

ルの Root-Mean-Square と，基本ベクトルとを要素毎に比較して，前者が大きければ文書ベクトルの新たな要素として置き換える．

$$d'_{j,i} \equiv \max(d_{j,i}, x_{j,i}), \quad (2)$$

$$x_{j,i} \equiv \sqrt{\frac{1}{m}\sum_{l=1}^{m} {k_{l,i}}^2} \quad (3)$$

ただし，$k_{1,i},\cdots,k_{m,i}$ は文書 d_jが属す文書群 $\mathcal{K}_1,\cdots,\mathcal{K}_m$の代表ベクトルの第 i 要素である．

2.4 検索システム R^2D^2

R^2D^2(RetRieval system for Digital Documents) は RS モデルを適用した文献検索システムで，NTCIR[*1] より提供を受けた国内の学会発表抄録データベース 332,921 件を対象に検索を行うものである．本研究では，NTCIR の方法に則り，「テストコレクション 1（予備版）」として用意された自然文 1 フレーズずつ計 30 件の問い合わせについて，それぞれ最大上位 1000 件におけるランク A 正解（完全にレレバント）の再現率と適合率を求めた．

3 query expansion

3.1 自動 query expansion 手法

query expansion は，検索者が入力した問い合わせ表現の関連語をシソーラスあるいは検索対象のデータベースから選択して問い合わせに加えることで，問い合わせの意味曖昧性に対処する手法である．検索者の意図と合致する語だけを自動的に選択して補うことは困難であるため，候補となる語を列挙するにとどめ，選択は検索者自身が行うという方式が一般的である．

関連語の抽出手法はシソーラスを用いた方式と予備検索から関連語を抽出する relevance feedback の方式とに大別されるが，前者は辞書構築のコストや問い合わせとの関連度を動的に決定することの難しさなどの課題を抱えており，後者は予備検索の精度が関連語の質を左右するというジレンマを持っている [1, 4].

また，relevance feedback によって補う語の数を数百以上に増やした類似文書検索が研究されており，特許検索やカテゴリが同じ文書を検索するといった，レレバント条件の緩い場合には成果をあげている [5, 6]．しかし問い合わせの意図がぼやけるため，レレバント条件が比較的厳しい検索には適さない．

Mitra らは relevance feedback の際に補う語の共起確率を考慮することで完全自動の query expansion の性能を向上させる手法を提案しており [3]，我々はそれを元に R^2D^2に完全自動の query expansion 機能を導入した．以下にその内容を示す．

tf·idf に基づく索引テーブルに対して元の問い合わせ表現で検索を行った結果の上位 D 件の文書に含まれる自立語を抽出し，新たな検索語 $\mathrm{t_{new}}$として問い合わせとの関連度 r を次式で求める．

$$r(\mathrm{t_{new}}) \equiv \frac{1}{|\mathcal{D}|}\sum_{\mathrm{d}\in\mathcal{D}} f(\mathrm{d})^2 \quad (4)$$

$f(\mathrm{d}) \equiv$ (文書 d に含まれる本来の検索語の数)^2

$\mathcal{D} \equiv$ (新しく加える語 $\mathrm{t_{new}}$を含む文書の集合)

次に，関連度の高いほうから T 語を検索語に補う．ただし新たな問い合わせ表現による各文書の得点は，

$$d \equiv \sum_{\mathrm{t}\in(\text{元の表現})} r(\mathrm{t}) + \sum_{i=1}^{T} r(\mathrm{t_{new}}_i) \times \min_{j=1}^{i-1}(1 - P(\mathrm{t_{new}}_i|(\text{元の検索語} \cup \mathrm{t_{new}}_j))) \quad (5)$$

とする．ただし $P(\mathrm{t_{new}}_i|(\text{元の検索語} \cup \mathrm{t_{new}}_j))$ は，加えた検索語 $\mathrm{t_{new}}_i$と，それよりも関連度の高い検索語の共起確率であり，以下の式で推定される．

$$\frac{\left(\begin{array}{c}\text{全文献の中で検索語 } \mathrm{t_{new}}_i\text{を含み，かつ} \\ \mathrm{t_{new}}_j\text{あるいは元の検索語のいずれかを含む文書数}\end{array}\right)}{(\text{全文献の中で検索語 } \mathrm{t_{new}}_i\text{を含む文書数})}$$

3.2 パラメータと性能の定性的関係

3.2.1 補う語の数

本手法における，補う語の数 T と性能への寄与の関係を調べた．

図 2 は，語を選ぶ文書数 D を 5 に固定した場合の $T = 10, 20$ それぞれの 11 点平均適合率である．10 語を補った場合には上位候補の適合率も全体を通しての適合率も改善がみられるが，20 語を補っ

[*1] NACSIS Test Collection for IR systems

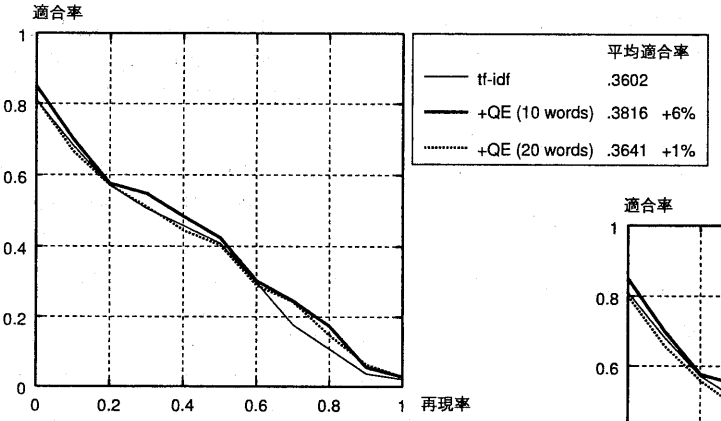

図 2 query expansion で補う語の数と検索精度の関係 (語を選ぶ文書数 D は 5 で固定)

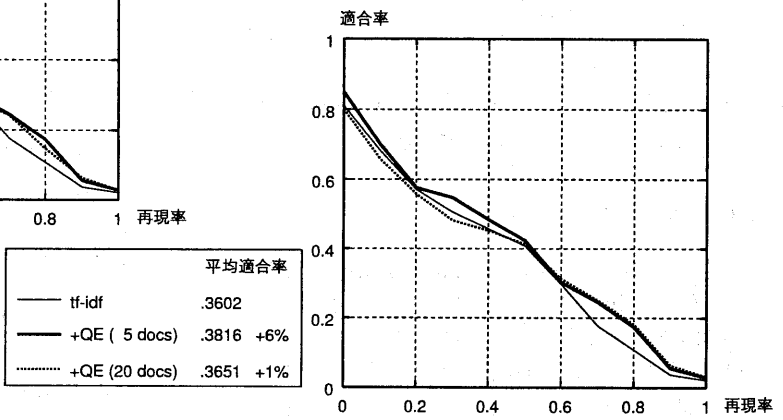

図 3 query expansion で語を選ぶ文書数と検索精度の関係 (補う語数 T は 10 で固定)

た場合には上位候補に対する効果が失われてしまっている．これは問い合わせが複数の話題を含んでいる時に，補った語が一部の話題にだけ偏り，その話題に関する点数だけが重視されてしまう傾向が強まるからである．対策として式 (5) で共起確率を考慮しているが，それも語数が増えると十分な効果を表すことができていない．

3.2.2 正解とみなす文書数

図 3 は補う語数 T を 10 に固定した場合の $D = 5, 20$ それぞれの 11 点平均適合率である．$D = 20$ では補う語数を増やした場合と同様，上位候補に対する精度向上の効果が失われている．これは，みなし正解を増やせばその中の不正解の割合が増加し，ノイズとなる語が補われやすくなるからと考えられる．

3.3 query expansion と RS モデルの比較

図 4 は，query expansion あるいは RS モデルを適用することによる平均適合率の向上あるいは劣化を 30 件の問い合わせのヒストグラムとして示したものである．図の左側に見るように query expansion は問い合わせごとの効果のばらつきが大きく，図の左下の領域に入っている問い合わせのように劣化してしまうことも少なくない．一方図の右側に対比した RS モデルは全般に変化 0 よりも上方に分布していることから，問い合わせによっては殆んど効果がない場合はあっても性能に悪影響を与えることは稀であるといえる．

3.4 自動 query expansion の問題点

以上の実験結果をまとめると，自動 query expansion は元の問い合わせ表現に対する検索結果の適合率が低い場合に十分な効果を上げることができず，みなし正解を多くとると不正解文書が含まれる率が増し，結果としてノイズとなる語が補われてしまうという問題を持っているといえる．

4 RS モデルと query expansion の融合

4.1 代表ベクトルに基づいた拡張語の選択

3 章の実験によって明らかになった自動 query expansion の問題を克服するものとして，我々は RS モデルにおける著者キーワードによる文書クラスタを用いて補う語を選択する手法を提案する．

すなわち，文書クラスタに付与された代表ベクトルは文書ベクトルと同次元であり，query ベクトルと代表ベクトルの間の関連度を求めることが可能で

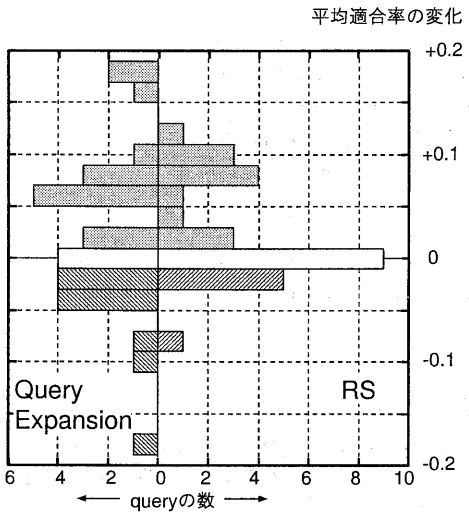

図 4　問い合わせごとの平均適合率の変化

あるという性質を利用して，問い合わせとの関連度の大きい C 個のクラスタに含まれる索引語から式 (4) による関連度の大きい順に T 語を選び，問い合わせに補う．

図 5 は RS モデルによる検索に query expansion を適用する際に，予備検索の上位候補 5 件から語を選んだ場合と，問い合わせとの関連度の高い 20 クラスタから選んだ場合における 30 件の問い合わせの 11 点平均適合率である．表 2 は同じ実験における問い合わせごとの平均適合率の変動を示したものである．各列は予備検索の上位候補 5 件中の正解件数である．

予備検索の上位 5 件から語を選ぶ場合，その中に正解が全く含まれていない，あるいは 1 件しか含まれていない場合，補われる語は検索者の意図を反映できておらず，結果として適合率が低下することが分かる．一方，提案手法では適合率がもともと低い場合にも精度を上げることに成功している．

これは次のように解釈できる．予備検索の上位候補というのは問い合わせごとに関連度の高い文書で動的にクラスタを作り，統計的に重要語を抽出する手法である．この際，統計的誤差を小さくするためには文書数を増やす必要があるが，文書数の増加は不正解文書の増加でもあるというジレンマを持っている．

これに対して文書関連性による静的なクラスタを用いた場合，各クラスタは十分な数の文書を含んで

表 1　問い合わせ「クラスタリングにおける特徴次元リダクション」の平均適合率

tf·idf	0.0716	
tf·idf+QE(上位 5 文書から)	0.0320	(–55%)
RS	0.0780	(+9%)
RS+QE(上位 5 文書から)	0.0302	(–58%)
RS+QE(20 クラスタから)	0.1508	(+111%)

※ query expansion で補った語数は全て 10.

いて，その代表ベクトルは一つの話題に関する重要語を統計的に抽出したものといえる．よって，問い合わせに関連のある代表ベクトルを組み合わせることで適切な関連語を選び出すことができる．

具体例として「クラスタリングにおける特徴次元リダクション」という問い合わせの検索結果を調べると，「リダクション」という表現は正解中には登場せず，「特徴抽出」と表現しているため，クラスタリングにだけ関係のある文書が上位候補に上がってしまい，上位 5 件中では 1 件だけが正解であった．この上位候補から抽出された「テクスチャ，画像，領域」などの関連語はますますクラスタリングに偏った文書の得点を押し上げるので，結果として精度を下げてしまっている．一方，提案手法では「多次元空間ベクトル」などのクラスタから「抽出，最小，縮小」などの関連語を補うことができ，検索精度の向上に成功している（表 1）．

ただし，予備検索が高精度であった場合には，その上位候補から選んだ重要語の正解弁別性はクラスタから抽出したものよりも高い．これは表 2 の高適合率部分において表れている．すなわち，高適合率の問い合わせは検索語の有無で単純に正解・不正解が分離しやすく，もともとの検索語以上に弁別性を持った語を静的なクラスタから選ぶことの困難さを示していると考えられる．

以上をまとめる．予備検索の上位文書から語を選ぶ方式では，query expansion による性能向上が求められる状況，すなわち適合率が低い場合ほど効果が小さく，補う語数を増やしたり，みなし正解を増やすことでは対処できない．一方，クラスタから語を選ぶ方式ならば適合率が低い状況において効果をあげることができる．

4.2　考察

提案手法は既存の自動 query expansion の弱点であった，適合率が低い状況における検索精度の向上

表 2　問い合わせごとの平均適合率の変化
–2/+1 とは，予備検索の上位 5 件の候補中 x 件が正解である問い合わせが 3 件あり，そのうち 1 件は query expansion によって精度が向上し，2 件は劣化したことを示す．

上位 5 件中の正解数	0	1	2	3	4	5	合計
上位候補文書から	–2/+0	–3/+1	**–1/+3**	**–2/+7**	**–2/+3**	–3/+3	**–13/+17**
代表ベクトルから	**–0/+2**	**–1/+3**	**–1/+3**	–6/+3	**–1/+4**	–4/+2	**–13/+17**

を達成した．これは RS モデルにおける代表ベクトルがそれぞれの話題の特徴的な索引語を抽出できていることを示す結果でもある．一方，高適合率の問い合わせに対しては既存手法で弁別性の高い語を選び出すことが分かった．実際の検索システムでは，予備検索の結果を元に手法を使い分けることが可能である．つまり，予備検索の結果中に検索者が正解と思う文書が少ない場合には提案手法を用い，正解が多い場合には既存の手法を用いて query expansion を行うようにすればよい．

5　おわりに

本論文は低適合率の問い合わせに対しても精度向上を果たす自動 query expansion として，RS モデルと融合した手法を提案し，実験によりその有効性を示した．

なお筆者らは，NACSIS コレクション (NTCIR) ワークショップに参加し，本研究では，NACSIS 研究開発部が「学会発表データベース」のデータの一部を使用して，データ提出学会[*2]の理解の下に構築した「テストコレクション 1（予備版）」を利用した．

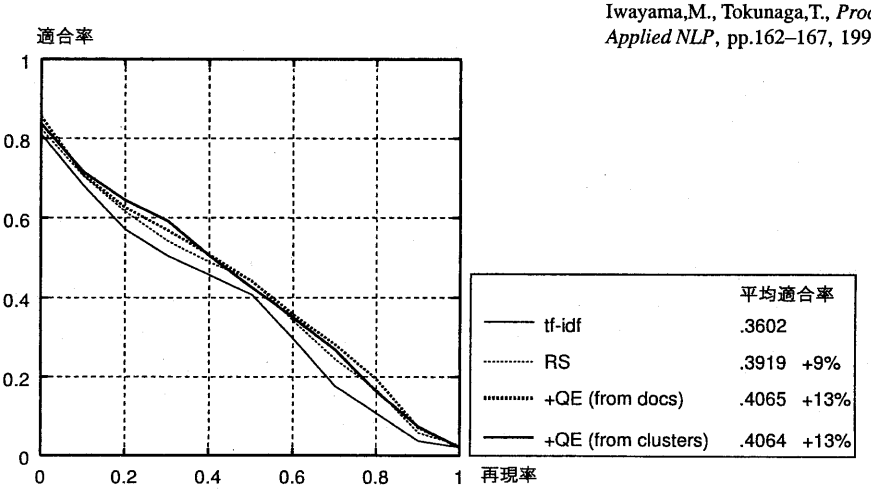

図 5　RS モデルに query expansion を組み合わせた場合の 11 点平均適合率

参考文献

[1] “文書関連性を考慮した検索方式,” 金沢 輝一, 高須 淳宏, 安達 淳, 情処研報, 98-DBS-116(2)-48, pp.165–172, Jul., 1998.

[2] “関連性の重ね合わせモデルによる文書検索,” 金沢 輝一, 高須 淳宏, 安達 淳, 信学会 第 10 回データ工学ワークショップ *DEWS'99*, 5B-5, 1999.

[3] “Improving Automatic Query Expansion,” Mitra,M., Singhal,A., Buckley,C., *SIGIR'98*, pp.206–214, 1998.

[4] “The impact of query structure and query expansion on retrieval performance,” Keäläinen,J., Järvelin,K., *SIGIR'98*, pp.130–137, 1998.

[5] “類似検索における単語寄与度に基づく重要語選択手法の検討,” 帆足 啓一郎, 青木 圭子, 松本 一則, 橋本 和夫, 情処学会第 57 回全国大会講演論文集, Vol.3, pp.239–240, Oct., 1998.

[6] “A Probabilistic Model for Text Categorization: Based on a Single Random Variable with Multiple Values,” Iwayama,M., Tokunaga,T., *Proc. of 4th Conference on Applied NLP*, pp.162–167, 1994.

[*2] http://www.rd.nacsis.ac.jp/~ntcadm/acknowledge/thanks1-ja.html 参照