データベースシステム　122－8
(2000. 7. 26)

英語テキストにおける関連性の重ね合わせモデルの検索特性

金沢 輝一† 高須 淳宏‡ 安達 淳‡

†東京大学大学院工学系研究科 ‡国立情報学研究所

〒101-8430 東京都千代田区一ツ橋 2-1-2 国立情報学研究所
TEL: 03–4212–2681 E-mail: {tkana, takasu, adachi}@nii.ac.jp

あらまし 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，著者キーワードなどの情報に基づいて文書をクラスタリングすることで，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ものである．筆者らはこれまでに情報検索システム評価のための大規模テストコレクション NTCIR-1 を用いて提案手法の評価を行い，有効性を示してきた．本稿ではテストコレクション TREC を用いた評価を行い，NTCIR-1 上での結果と比較することによって，言語，文書の種類，問い合わせ表現などの違いが提案手法の検索特性にどのような影響を及ぼすかを検証する．

キーワード 情報検索，ベクトル空間モデル，文書ベクトル修正, RS モデル，TREC, NTCIR

Retrieval Effectiveness for English Text Using the Relevance-based Superimposition Model

Teruhito KANAZAWA† Atsuhiro TAKASU‡ Jun ADACHI‡

† Graduate School of Engineering, University of Tokyo

‡ NII (National Institute of Informatics)

NII, 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, JAPAN

TEL: +81–3–4212–2681 E-mail: {tkana, takasu, adachi}@nii.ac.jp

Abstract We have proposed a Relevance-based Superimposition (RS) model to solve the problems of semantic ambiguity on information retrieval. This method partitions the documents so that the relevant documents fall into the same cluster based on the keywords given by authors, and it enables more accurate estimation of the weights of index terms than conventional methods such as tf·idf. We evaluated our method and showed the effectiveness of it using the NTCIR-1, which is a large-scale test collection for information retrieval. In this paper, we evaluate the proposed method using the TREC test collection, and examine the relation of the retrieval effectiveness of the proposed method to the kind of language and the characteristic of documents and queries.

key words information retrieval, vector space model, document vector modification, RS model, TREC, NTCIR

1 はじめに

計算機の性能向上と普及に伴って電子化文書の流通・蓄積量は驚くべき速度で増加している．情報の効率的な利用には，必要な情報を選び出すための検索技術が不可欠であるが，従来のキーワード入力型の検索手法では自然言語の持つ曖昧性によって検索精度が低下することが指摘されていた．

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の向上を図る手法を提案し，情報検索システム評価のための大規模テストコレクション NTCIR-1 上での評価を行ってきた [1]．NTCIR-1 は学術論文の抄録データベースから構成されており，日本語の問い合わせに対して日本語あるいは英語の文書を検索する場合の精度を評価することができる [2]．一方，英語の大規模テストコレクションとしては TREC が存在しており，記述言語はもとより構成文書の種類，問い合わせの特徴においても NTCIR-1 とは大きく異なっている．本稿では TREC 上で提案手法の評価を行い，NTCIR-1 上で行った評価と比較することで，先に述べたようなデータベースや問い合わせの特徴と提案手法の検索性能との関係について考察を行う．

次章では筆者らが提案している関連性の重ね合わせモデルの定義を述べ，3 章で実装システムについて説明する．4 章で評価実験の方法と結果を示し，最後に 5 章で実験結果に基づいて言語，文書，問い合わせの特徴が提案手法の検索精度に与える影響について考察を行う．

2 RS モデル

本節では筆者らが提案している関連性の重ね合わせモデル (RS モデル) の概要を述べる．RS モデルは検索対象文書の意味曖昧性に対処することで検索精度の向上を図った手法であり，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の特徴ベクトルを拡張するというものである．

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．今回の実験ではクラスタは文書から抽出したキーワードによって形成されている．例えばデータベース中にキーワード A と B の 2 つのキーワードが存在した場合，キーワード A を含む文書はクラスタ C_Aに，キーワード B を含む文書はクラスタ C_Bに属する．また，キーワード A, B をともに含む文書は C_Aと C_Bの両方に属するものとする．

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修正の 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．α-関数族 [3] から派生する幾つかの関数の評価 [4] 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 r_iを次のように求める関数である．

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

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．

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

$$x_{j,i} \equiv \sqrt{\frac{1}{m}\sum_{l=1}^{m} {r_{l,i}}^2} \tag{3}$$

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

3 情報検索システム R^2D^2

筆者らは文献検索のためのシステム R^2D^2 (RetRieval system for Digital Documents) を作成し，こ

れにRSモデルを適用して評価を行っている．実装の詳細は文献[1]に述べられており，本論文では英語に対応する際の変更点のみを紹介する．

3.1 英語の形態素解析

空白などのデリミタによって単語を抽出し，冠詞，前置詞，代名詞などのストップワードを取り除いた後，Porterの語幹切り出しアルゴリズム[5]と不規則変化動詞の辞書を併用して語幹のレベルで索引を作成した．今回の実験では名詞句などの認識は行わなかった．

3.2 語の重み付け

R^2D^2では次に示す3つの統計量に基づいて語の重み付けを行っている．

- 文書中での語の出現頻度 (term frequency): f_T
- 全文書中で語を含む文書の数 (document frequency): f_D
- 語の共起頻度 (term cooccurrence): f_C

検索語 $\{q_0, q_1, ..., q_m\}$ から成る問い合わせ $\boldsymbol{Q}$ に対する文書 d_j の索引語 t_i の重みを，

$$w(i,j,\boldsymbol{Q}) \equiv f_T(j,i) \cdot f_D(i) \cdot f_C(i,\boldsymbol{Q}) \qquad (4)$$

と定義し，以下にそれぞれの改善の経緯を紹介する．

3.3 Term Frequency に基づく重み付け

NTCIRにおける予備実験で筆者らは一般的なtf·idfを含めた幾つかの関数を索引語の重み付けに適用し，それらの特性評価を行った．その結果，

$$f_T(j,i) \equiv \frac{1}{\pi}\arctan(tf_{j,i}) + 0.5 \qquad (5)$$

という式を用いた場合が一般的なtf·idfの式の

$$f_T(j,i) \equiv tf_{j,i} \qquad (6)$$

よりも検索精度を向上させることが判明した．ここで，$tf_{j,i}$ は索引語 t_i が文書 d_j に出現する回数である．

式(5)の特徴は図1のように $tf \to \infty$ に対して有界であるという点である．問い合わせが複数の語で構成されていた場合に式(6)では頻出の一語の重みが文書の得点に大きく影響してしまうが，式(5)では多くの種類の検索語が出現する文書の得点が高くなる．NTCIR-1は論文抄録であり文書長が比較的短く，重要語が文中に一回しか出現しない場合もあるので tf と重要度との相関は小さいと予想でき，予備実験によってこの事が確かめられたといえる．

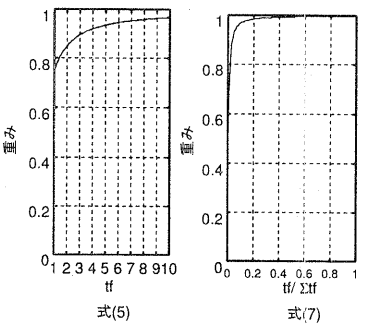

図1 tf に基づく重み付け

一方TRECはFederal Registerなど比較的長い文書も含まれており，文書長の偏差が大きいデータベースであって，文書長による tf の正規化が検索精度の向上にも効果的であると予想された．そこで式(5)を，

$$f_t(j,i) \equiv \frac{1}{\pi}\arctan\left(\alpha\frac{tf_{j,i}}{F(j)} + \beta\right) + 0.5 \qquad (7)$$

として，$\alpha, \beta, F(j)$ の最適化を行ったところ，TREC3全体に対しての最適値は $\alpha = 100, \beta = -0.5, F(j) = \sum_i tf_{j,i}$ であった．

3.4 Document Frequency に基づく重み付け

索引語 t_i を含む文書数を df_i，全文書の数を N としたときの $f_D(i)$ として一般的に用いられている

$$f_D(i) \equiv \log(N/df_i) \qquad (8)$$

と，NTCIR上での実験時に提案した，

$$f_D(i) \equiv \frac{2}{\pi}\arctan(N/df_i) \qquad (9)$$

を比較して，前者を用いた場合が検索精度が高いという結果が得られた．

3.5 Term Cooccurrence に基づく重み付け

NTCIRでは，文書 d_j に出現する検索語の種類を c_j，検索語 t_i が出現する文書の集合を Δ_i として，

$$f_C(i,\boldsymbol{Q}) \equiv \sqrt{\frac{1}{df_i}\sum_{\mathrm{d}_j \in \Delta_i}(c_j - 1)^2} \qquad (10)$$

を用いた．TREC では

$$c(i) \equiv \sum_{\mathrm{d}_j \in \Delta_i} \sum_{\mathrm{t}_k \in \mathrm{d}_j} f_D(k) \quad (11)$$

$$\overline{c}(i) \equiv \sum_{\mathrm{d}_j \notin \Delta_i} \sum_{\mathrm{t}_k \in \mathrm{d}_j} f_D(k) \quad (12)$$

$$f_C(i, Q) \equiv \log \frac{c(i)}{df_i} - \log \frac{\overline{c}(i)}{N - df_i} \quad (13)$$

を用いた．式 (10) では問い合わせの話題の中心ではないがデータベース全体で共起頻度の高い検索語があると，その重みが高くなってしまう傾向があった．例を表 1 に示す．この例では research と study という共起頻度の高い検索語の重みを式 (10) によって求めた場合，話題を表現している語の一つである treatment と同程度となっている．この問題を解消するために，式 (13) では，問い合わせの話題に関連度の高い文書集合における情報量 $\log \frac{c(i)}{df_i}$ と，補集合における情報量 $\log \frac{\overline{c}(i)}{N-df_i}$ との差分をとっている．

4 評価実験

4.1 NTCIR-1

NTCIR(NACSIS Test Collection for Information Retrieval) は学術情報センター[*1]による文書検索システム評価に関するプロジェクトで，1999 年に第一回ワークショップが開催された [2]．評価用コーパス NTCIR-1 は国内の学会発表の抄録データベースから構成されており，日本語を基本言語としているが約半数の文書には英語対訳が存在し，言語横断システムの評価にも利用可能となっている．随時検索評価用の問い合わせは日本語で 83 件が用意されている．

NTCIR-1 の文書には全体で統一された構造・要素タグが付与されており，索引語の抽出はタイトル，要旨，著者が付与した自由キーワードより行い，RS モデルのクラスタリングは著者キーワードに基づいて，同じ著者キーワードを含む文書の集合を一つのクラスタとした．また，問い合わせは「検索要求」要素のみを用いた．

表 1　ストップフレーズを含む問合わせに対する重み付け
問い合わせ "Documents will focus on studies as to causes of multiple sclerosis and on research efforts to develop treatments and/or cures for it."
値は最大値に対する比率に正規化してある．
式 (10) では treatment の重要度は research や studies などの一般語と同程度となるが，式 (13) ではそれら一般語の重要度は小さく，treatment との差が明確になっている．

検索語	式 (10)	式 (13)
sclerosis	1.0	1.0
cures	0.83	0.59
research	0.59	0.21
treatment	0.57	0.31
studies	0.55	0.16

[*1]現在は国立情報学研究所

4.2 TREC

TREC(Text REtrieval Conference) は NIST[*2]と ISTO/DARPA[*3]がスポンサーとなって 1992 年から行われている文書検索システムに関するワークショップの名称である [8]．評価用コーパスは Associated Press newswire，Wall Street Journal などの報道記事，Federal Register などの行政文書を含む英語の文書群から構成されており，年度ごとに 50 件の随時検索評価用の問い合わせと検索対象の文書群が用意されている．これらはワークショップに参加しない場合でも研究目的で使用することが可能となっており，またワークショップに参加したシステムの検索精度の概要が公開されているので，英語に対する検索システムの標準的な評価方法として利用することができる．

TREC の文書は AP，WSJ，FR などの文書群ごとに異なる構造・要素タグ付けがなされており，NTCIR のようにデータベースに付与されている情報に基づいて全文書を対象にしてのクラスタリングを行うことはできない．今回の実験では NTCIR に対して適用した手法との比較を容易にするために TREC コーパスを文書種ごとに分割し，NTCIR の著者キーワードに相当する情報が抽出できると判断した AP，SJM に限定して評価を行った．この際，部分セットに正解が 5 文書未満しか含まれていない問合わせは平均適合率の計算対象から除外した．

4.2.1 TREC3 AP

Associated Press newswire の文書群の主な構成要素は次の通りである．

[*2]National Institute of Standards and Technology,
米国標準技術局

[*3]Intelligent System Technology Office of the Defense Advanced Research Projects Agency,
米国 国防総省 高等研究計画局 情報技術部

- FIRST, SECOND … 概要キーワード．FIRSTはSECONDをさらに要約したもの．
- HEAD … ヘッドライン．
- TEXT … 本文．

索引語の抽出はSECOND, HEAD, TEXTより行い，RSモデルのクラスタリングはSECOND要素とHEAD要素から抽出した索引語に基づいて，同じ索引語を含む文書の集合を一つのクラスタとした．

TREC3の問い合わせはtitle, desc, narrの3要素で構成されているが，今回の実験ではNTCIRの実験との比較のために，descのみから検索語を抽出した．

4.2.2 TREC4 AP

AP文書群の構成要素はTREC3とTREC4で同一であったので，索引語の抽出方法もTREC3と同様とした．TREC4の問い合わせはdescのみで構成されており，全体から検索語を抽出した．

4.2.3 TREC4 SJM

San Jose Mercury Newsの文書群の主な構成要素は次の通りである．

- SECTION … 新聞の「面」．スポーツ，生活など．
- HEADLINE … ヘッドライン．
- LEADPARA … リードパラグラフ．
- DESCRIPTION … 人手で割り振られたカテゴリキーワード．
- MEMO … 一行程度の概要．
- TEXT … 本文．
- CITY … 話題の中心となる都市．
- CAPTION … 図表のキャプション．

索引語の抽出はHEADLINE, LEADPARA, DESCRIPTION, TEXT, CAPTIONより行い，RSモデルのクラスタリングは，DESCRIPTION要素に基づいて，同じキーワードを含む文書集合を一つのクラスタとした．

4.3 結果

表2に，それぞれのテストセットにおける検索精度を示す．まず，抽出された文書クラスタの数がテストセット間で大きく異なった．これはNTCIR-1の著者キーワードは自由キーワード，TREC APのキーワードはヘッドライン，TREC 4 SJMのキーワードは統制キーワードというそれぞれの性質を反映している．次にbaselineの平均適合率を基準にRSモデルの効果をみると，NTCIR-1では9〜12%であったのに対し，TREC APでは3〜4%，TREC SJMで6%とやや向上率が小さくなっている．問い合わせ別の効果をみても，NTCIR-1では全体の3割以上の問合わせで平均適合率が0.05以上向上したのに対し，TRECでは5〜20%の問い合わせに留まっている．一方，平均適合率が0.05以上低下した問い合わせの数では，NTCIRの1〜2件に対して1〜5件と増加している．

図2は11点平均適合率のグラフである．NTCIR-1では全域で適合率が向上しているが，TRECでは再現率0，すなわち最上位候補で適合率の低下が発生している．また，適合率の向上の度合いもNTCIR-1の場合に比べて小さいことが分かる．

表3，4はそれぞれRSモデルの効果が現れた問い合わせと，そうでない問い合わせの例である．表3ではdeath, penaltyなどのクラスタが作用してcapital punishmentという検索語が補われた正解文書が順位を上げ，相対的に不正解文書は順位を下げている．一方，表4では正解文書がfuelというクラスタに属していながら，このクラスタ内の文書でgasolineあるいはoilといった問い合わせと同一の表現がほとんど用いられていないために順位を上げることができず，より広い分野を表すenergyなどのクラスタに属する文書が順位を上げ，その中には不正解文書も多く含まれていることが分かる．

5 考察

NTCIRコーパス上での実験と比較すると，TRECコーパスでのRSモデルの寄与は小さいことが分かる．その原因と思われるテストセット間の特徴の違いを以下に考察する．

5.1 文書の特性

NTCIRコーパスが学術文書であるのに対して今回の実験で用いたTRECコーパスのサブセットは報道記事である．前者には，例えば「無線通信」という概念を表現する場合に著者によって「無線チャネル」あるいは「モバイル」といった別の語を用いる場合があり，これらの表現の差異をRSモデルの文書ベクトル修正作用が吸収して検索精度を高めている．一方，報道記事では同じ概念を表現する語は

表 2　テストセットの特徴と検索精度

文書群	NTCIR-1		TREC 3 AP	TREC 4 AP	TREC 4 SJM
文書数	339,483		164,597	158,240	90,257
クラスタ作成の情報源	著者キーワード		HEAD と SECOND 要素		DESCRIPTION 要素
有効クラスタ数	77,996		10,671	10,341	776
問い合わせ数	30 (予備用)	53 (提出用)	49	45	37
baseline の平均適合率	.3602	.2933	.2865	.2466	.2343
RS モデルの平均適合率	.3919	.3289	.2944	.2571	.2476
向上率	+.0317, 9%	+.0356, 12%	+.0079, 3%	+.0105, 4%	+.0133, 6%
平均適合率が 0.05 以上向上した問い合わせの数と割合	9 30%	18 34%	5 10%	9 20%	2 5%
平均適合率が 0.05 以上低下した問い合わせの数と割合	1 3%	2 4%	3 6%	5 11%	1 3%

図 2　11 点平均適合率

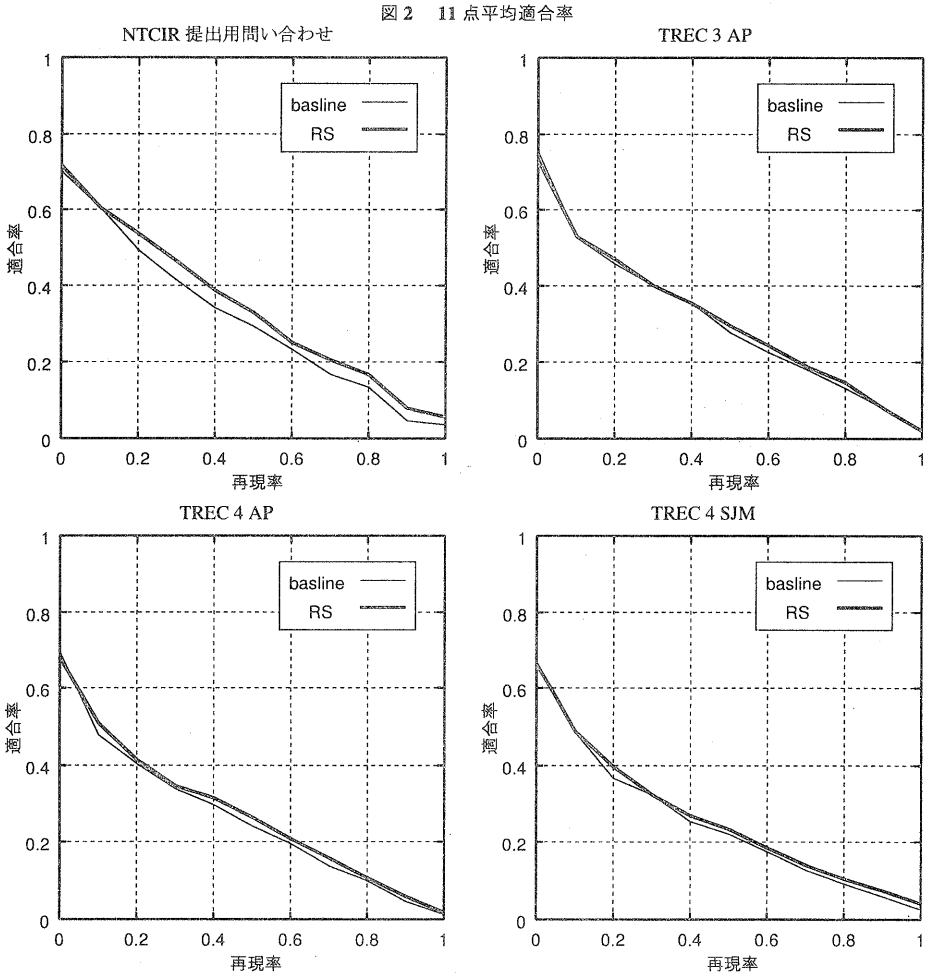

表3　RSモデルによる検索精度向上の例
問い合わせ #222：Is there data available to suggest that capital punishment is a deterrent to crime?
順位の変化は「basline → RSモデル」

文書番号	題	順位の変化	キーワード
正解			
AP900405-0062	Government Proposes Outlawing Capital Punishment	13 → 6	ireland, death, penalty, govern, propose, outlaw, capital, punish
AP880708-0047	Parliament Approves Death Penalty For Drugs, Trafficking In Women	63 → 34	bangladesh, death, parliament, approve, penalty, drug, traffic, woman
不正解			
AP881014-0121	Senate Might Finish $2.6 Billion Drug Bill Today	5 → 29	congress, drug, senate, finish, billion, drug, bill, today
AP900402-0048	Police Smash Computer-Hacking Operation	30 → 393	australia, hack, police, smash, computer, operate

表4　RSモデルによる検索精度低下の例
問い合わせ #237：Identify alternative sources of energy for automobiles. Include additives to gasoline that either decrease pollution or reduce oil consumption.
順位の変化は「basline → RSモデル」

文書番号	題	順位の変化	キーワード
正解			
AP881014-0188	Reagan Signs Bill To Encourage Alternative Fuels in Autos	19 → 53	reagan, alternate, fuel, sign, bill, encourage, auto
AP900213-0058	Administration Putting Brakes on Alternative Fuel Cars Idea	20 → 51	clean, air, administr, put, brake, alternate, fuel, car, idea
不正解			
AP900802-0205	The Energy Problem Drifts Into The Background Also moved in advance	41 → 16	busy, mirror, energy, problem, drift,background, move, advance
AP900214-0207	Eastern Europe Economies At-a-Glance With BC-EUR—Economic Challenge	64 → 17	eur, glance, eastern, europe, economy, glance, bc, eur, challenge

意識的に統制される傾向にあり，文書ベクトル修正の寄与は学術文書と比べて小さいと推測される．

NTCIR コーパスの著者キーワードに比べて，TREC の概要キーワードは種類が少なく，特に SJM ではわずか 776 語しか有効なキーワードが存在しなかった．中には話題を表していない “Digest Briefs” のようなキーワードのみが付与されている文書もあるので，TREC のキーワード要素に基づいて作成した文書クラスタには話題と適切に対応していないものもある．また，一つの文書に付与されている数が NTCIR の 4〜8 語に比べて 1〜3 語と少なく，クラスタに全く属さず文書ベクトル修正の作用を受けない文書が多数存在するという問題もある．今後は付与キーワード以外の情報に基づくクラスタリング手法の導入を検討する必要がある．

5.2　問い合わせの特性

NTCIR の問い合わせは学術用語を含むものが多く，問い合わせ中の重要語が無関係な文書に頻出することは稀である．一方 TREC の問い合わせは “What are the trends and developments in retirement communities?” のように単語レベルでは話題を象徴できないような場合が多い．この問題に対処するためにフレーズレベルでのマッチングを導入した上で再度評価を行いたい．

また，RSモデルは文書データベース側の表現の差異を吸収する働きを持っているが，問い合わせに一般的でない表現を用いている場合には query expansion などの手法が有効である．例えば “What is the extent of U.S. arms exports?” では兵器という意味で arms を用いているが，コーパス中では arms をこの意味で用いることは稀であり，代りに weapon, military などの表現が用いられている．このような場合は問い合わせの表現をデータベースに合わせて補正した上で RS モデルを適用することで一層の効果が得られると予想される．

5.3 言語依存性

NTCIR，TREC コーパス上での実験を通して，形態素解析の精度や重み付けのパラメータが検索精度のベースラインに大きく影響することが分かった．一方，RS モデルでは代表ベクトル生成関数，文書ベクトル修正関数として Root-Mean-Square が最適であるという結果は NTCIR-1，TREC とも変わらなかった（表 5）．言語などの特性が大きく異なるデータベース間で最適条件が一致したことから，複数データベースの検索結果の統合や言語横断検索のように他手法ではパラメータの最適化が難しいとされる分野においてもチューニングのコストをかけることなく効果を発揮できるものと期待される．

6 結論

本論文では関連性の重ね合わせモデルに基づく検索手法を日本語と英語の大規模テストセット上で評価することにより，その特性を明らかにした．提案手法は主に文書データベース側の同概念異表記の問題に対して有効であり，このような特徴を持つデータベースとして学術文書などが存在することを示した．一方で新聞記事データベースでは文書の同概念異表記よりも問い合わせ表現の曖昧性の問題が大きく，フレーズマッチング，query expansion などの手法と提案手法との融合による精度向上の可能性を指摘した．

ここで WWW の文書に注目すると，文書の特徴としては作成者が各々の表現を用いているために同概念異表記の問題を持っており，提案手法による検索精度の改善が期待できる．しかし WWW 文書の大部分はクラスタリングに有効なキーワードが予め付与されているわけではないので，それ以外の情報に基づくクラスタリングを検討する必要がある．WWW 文書検索の分野ではリンク情報に基づいたクラスタリング [6] や，情報抽出によって半構造データ化する技術 [7] などが提案されており，今後はこのような手法の導入を検討していきたい．

表 5　RS モデルのベクトル関数と検索精度

関数	NTCIR-1	TREC 4 SJM
RR	**.4024 (+7%)**	**.2476 (+6%)**
RA	.4003 (+6%)	.2467 (+5%)
AR	.3947 (+5%)	.2451 (+4%)
AA	.3920 (+4%)	.2438 (+4%)
baseline	.3772	.2343

関数の記号は 1 文字目が代表ベクトル生成関数，2 文字目が文書ベクトル修正関数で，R は Root-Mean-Square，A は算術平均を意味する．

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[*4]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．

参考文献

[1] Kanazawa, T., "R^2D^2 at NTCIR: Using the Relevance-based Superimposition Model," *Proc. of NTCIR Workshop 1*, pp.83 – 88, Aug. 1999.

[2] NTCIR: http://www.rd.nacsis.ac.jp/~ntcadm/

[3] 林 幸雄, "個人選考による情報アクセスに適したデータモデルについて," 情処研報 98-DBS-116(2), pp.381 – 388, July 1998.

[4] 金沢 輝一, 高須 淳宏, 安達 淳, "関連性の重ね合わせモデルによる文書検索," 電子情報通信学会第 10 回データ工学ワークショップ (DEWS'99), Mar. 1999.

[5] Baeza-Yates, R., and Ribeiro-Neto, B., "Modern Information Retrieval," Addision-Wesley, 1999.

[6] 小林 伸行, 北川 文夫, "WWW 上のページセットの抽出とそれを用いた検索," 電子情報通信学会第 10 回データ工学ワークショップ (DEWS'99), Mar. 1999.

[7] 山田 洋志, 福島 俊一, 松田 勝志, "Web ページからのタイプ別情報抽出・分類方式," 情処研報 00-FI-57, pp.143 – 150, Mar. 2000.

[8] TREC: http://trec.nist.gov/

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