データベースシステム 116－48
（1998. 7. 9）

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‡

† 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

情報検索においては自然言語の意味曖昧性が大きな問題となっており，ベクトル空間モデル上で問合せ表現のベクトルを拡張する query expansion などの手法が存在する．しかし問合せ表現は情報量が小さいために，検索者の意図を汲み取って的確にベクトルを拡張することは難しい．著者らは文書関連性を用いて文書ベクトルを拡張することで検索性能を向上させる手法を提案する．提案手法では検索テーブル作成時に文書関連性に基づいて文書集合を作り，この集合を単位として補う要素を決定することで精度の向上を図る．評価実験として，学術論文に予め付与されているキーワードを情報源として文書間の関連性を抽出，この関連性を元にベクトルを補って，tf·idf モデルや query expansion との性能比較を行う．

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，文書関連性

A Retrieval Method Based on Relevance of Documents

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

Ambiguity of meaning is a serious problem in information retrieval, and query expansion in the vector space model is one of the typical methods, which expands the query vectors to cope with this problem. However, queries tend to have less information for fitting query vectors to the latent semantics, which are difficult to express in a few query words given by users. We propose a document expansion method which expands the document vectors based on relevance of documents. The proposed method, in which document sets are prepared based on the relevance of documents at the time search table is constructed by adding words for each set, is expected to increase the query precision. In this paper, we evaluate our method through retrieval experiments in which the relevance of documents extracted from scientific papers, and the comparison with tf·idf and query expansion methods is described.

Keyword information retrieval, vector space model,
document vector expansion, relevance of documents

1 はじめに

情報検索において，問い合わせ表現 (query) をシステム側で補うことで検索性能を向上させることができる．代表的な手法として query expansion [1] が挙げられるが，元々情報量の小さい query を高い精度で拡張することは難しく，補われた語彙によって不要な文書が検索される率が高まることがある．

本論文ではデータベース内の文書関連性を利用して検索可能な語を文書単位に補うことで性能を向上する手法を提案する．

次章では自然言語の意味的曖昧性が情報検索に及ぼす影響と従来の対処法を紹介し，3 章で文書関連性を考慮したベクトルモデルを提唱する．4 章では評価実験の結果を報告する．各文書の著者が付与したキーワードに基いて文書集合を作り，この集合単位でベクトルを拡張するモデルを設定して評価実験を行ったところ良好な結果が得られた．最後に問題点を整理，今後の研究方針について述べる．

2 意味的曖昧性の問題

文書検索処理は，検索者の意図と検索対象となる文書群を照合して適合するものを探すことである．問合せ表現 (query) として自然文あるいは語の列挙の形で入力を行う方法が一般的だが，query は検索者の意図を代表する表現の一つに過ぎず，表現とそれによって示される概念は一対一対応とは限らないために，query から検索者の意図を正確に汲み取ることが難しい場合もある（図 2(a)）．これは意味的曖昧性の問題と呼ばれており，情報検索においては実用的な検索精度を得るために曖昧性への対策が必要となる．

情報検索における意味的曖昧性の問題は，自然言語処理におけるそれとは異なる性質がある．自然言語処理では「前後の文章から情報を得て，表現の多義性を解消する」という，人間の文脈理解に相当する処理を目的としているのに対し，情報検索では query という非常に限られた情報を元に処理を行わなければならない．query の曖昧性は人間でも一意に解消できない場合が多いので，自然言語処理技術を応用しても完全に解消できるわけではない．

現在情報検索においては，以下に紹介するような手法が用いられている．

2.1 query expansion

query expansion は，query の表現に対応する概念の範囲を広く設定するものというイメージである（図 2(b)）．

2.1.1 基本的な手法

1. 予め静的な解析によって全ての表現と概念の関連度を計算し，概念 c_j との関連度を第 j 要素とするベクトルを表現ごとに作成する．表現 t_i のベクトルは

$$t_i \equiv \sum f(i,j) \cdot e_j \quad (1)$$

(e_jは j 軸の単位ベクトル)

$$f(i,j) \equiv \begin{cases} rel(i,j) & (rel(i,j) \geq \alpha) \\ 0 & (rel(i,j) < \alpha) \end{cases} \quad (2)$$

($rel(i,j)$ は表現 t_iと概念 c_jの関連度，αは閾値)

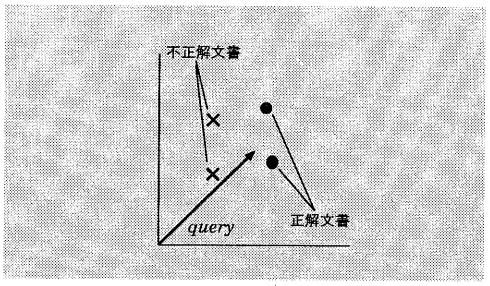
2. query が表現の集合 $\{t_1, ..., t_n\}$ で与えられた時，query ベクトルは t の和で表される．

表現と概念の対応は文書母集団から得ることが多いが，辞書から抽出することもある [2].

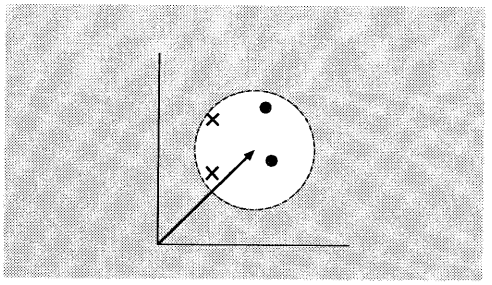
2.1.2 みなし正解中の重要語を用いる手法 [1]

1. 全ての文書に対して query との関連度を計算し，上位から一定数（あるいは割合）を「みなし正解」とする．
2. みなし正解中の重要表現（頻出語など）を抽出して，それらを query に補う．

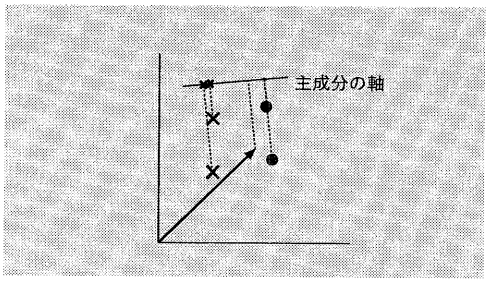
この手法では，検索の度にみなし正解群の重要表現を抽出する処理が必要だが，検索対象の文書母集団の変化に伴って表現と概念の関係も変化するような状況に動的に対応できるという利点を持つ．反面，みなし正解が検索者の要求に適合しているとは限らず，適切な表現を補えない可能性もある．



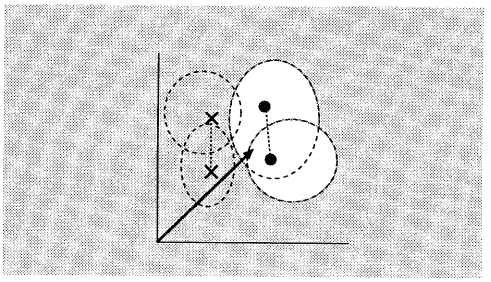
(a) 元の状態



(b) query expansion



(c) 意味空間の縮退



(d) 文書ベクトルの拡張（提案手法）

■ 図 1　意味的曖昧性の対策手法イメージ

提案手法は文書関連性を考慮して各文書のベクトルを拡張することでマッチングの精度を高めることを狙っている.

2.2　意味空間を縮退するモデル

意味空間を縮退することでも query expansion と同様に意味曖昧性の対策となる（図 2(c)）. また, 検索テーブルが小さくなるので処理コストの低減につながる.

2.2.1　Latent Semantic Index モデル

意味空間の多変量解析によって縮退を行うのが Latent Semantic Index モデルである [3].

LSI ではまず, 語 t_jの出現頻度を j 番目の要素とする文書ベクトルを用意する. 語の種類を n_tとすると各文書の意味ベクトルは n_t次元のベクトルとなる. 相関として文書単位での語の共起頻度を用い, 主成分分析によって情報量を余り損なわずに n_tよりも低次元のベクトルに縮退する.

2.2.2　意味の数学モデル

LSI では, 検索対象の文書母集団にとって情報量の小さい概念に対応する次元が縮退される. 概念の情報量は検索時に動的に変化するものであるため, 検索時に空間のスケール変換を行うという手法が考案されたが, その効果を十分に得るためには情報量の基準を文書母集団よりもっと普遍的なものにしたほうがよい. この問題に対し, 辞書を用いる手法を示したのが意味の数学モデル (Mathematical Model of Meaning) である [4].

MMM で問題となるであろう点は, 辞書の語彙集合が文書のそれを完全に包摂しているとは限らないことである. 専門用語が重要語となる論文データベースや新しい表現が登場する時事ニュース・WWW 等では, 辞書未定義語の扱いが性能を大きく左右するため, さらなる工夫が必要である.

■ 表 1　評価実験に用いた query と, 対応する正解文書の数

No.	フレーズ	正解文書数 A	正解文書数 B
1	ATM 網におけるトラヒック制御	9	43
2	類似図形パターン検索	5	35
3	WWW 情報空間でのユーザ支援	34	9
4	ニューラルネットを利用した画像の特徴量抽出	14	129
5	並列処理を用いた FFT の高速化	8	72

3　文書関連性を考慮した検索手法の提案

3.1　文書関連性の考慮とは

前章で紹介した意味的曖昧性への対策手法のうち, query expansion では, query という元々情報量の小さい要素に対する処理であるために, 再現率が向上する一方で適合率が犠牲になるという問題が生じる. すなわち前後の文章などから曖昧性を解消することができないのである. そこで, query expansion とは対照的に, 検索対象の文書側でベクトルを拡張する手法を考える（図 2(d)）. この場合,

- 文書ごとに文脈を解析してベクトルを拡張する.
- 文書間にもさまざまな関連性が存在するので, 関連する文書で部分集合を作り, これを解析することで文書単位の解析からは得られない情報を得る.

という 2 つの方向性が考えられるが, 今回は後者によって文書ベクトルを拡張する手法を提案する.

3.2　従来の手法との比較

query expansion と比較した場合, query という情報量の小さい要素を拡張するよりもベクトル拡張の精度が高まるので, 適合率の低下を抑制しながら再現率を向上させることができると予想される（図 2(d)）.

また, 意味空間を縮退する LSI や MMM では縮退によって「検索者にとっては異なる意図を表す query がシステムでは同じものとして扱われてしまう」可能性があるが, 提案手法ではこのような問題は生じない.

その反面, 単純な実装では検索テーブルが数倍から数十倍に膨らんでしまうという問題点を持っている（検索速度は query expansion と同程度と予想される）. これに関してはインデクスの構造などを工夫することで対処することが考えられる.

4　評価実験

提案手法によるベクトル拡張の効果を確認するために実験を行った.

4.1　検索対象データ

- NACSIS-IR の学会発表データベースより, 情報処理学会のコンピュータ関係の発表 13461 件分を利用. このデータベースは各発表に関して
  ・発表題目
  ・要約文
  ・発表者が付与した自由なキーワード（複数）
  が記録されており, 今回の実験では題目と要約文を検索対象とし（以下「文書」と呼ぶ）, 付与キーワードが完全一致するものを関連性がある文書とみなすモデルを用いる.
- データベースには誤字や表記の揺れ（外来語の長音記号や略称など）が存在するので, 前処理による表記の統一と誤字の修正を行い, 実験への影響を抑えるようにした.
- 題目と要約文は表記の修正後に形態素解析プログラム ChaSen[5] による解析を行い, 体言・用言のみを検索対象語として抽出した（抽出された語は 26095 種類）.
- 今回の実験では文書ベクトルは検索対象語 t_j の出現頻度を j 番目の要素とした. 以下, ここで求めたベクトルを基本ベクトルと呼び, 文書 d_n の基本ベクトルの j 番目の要素を d_j^n と表す.

- 検索要求と正解集合は松村らの研究[6]で同じ文書集合に対して作成された5件を使った. queryのフレーズと対応する正解文書数を表1に示す. 正解は,

  ランク **A** … 問い合わせの内容をほとんど全て含んだもの.

  ランク **B** … 問い合わせの内容の一部を含むなど, 関係の深いもの.

  という基準で定義されている.

4.2 文書ベクトルの拡張

1. 今回の実験ではキーワードは形態素解析を行わずに, 完全一致したものだけを同一キーワードとみなし, 同じキーワードが付与されていることをもって文書が関連性を持っていると判断する. データベース中にはキーワードが19853種存在し, 複数の文書に付与されていたものは5876種であった.
2. キーワード k_l が付与されている文書群 $\mathcal{K}_l$ の代表ベクトルを作る. 代表ベクトルの要素は $\mathcal{K}_l$ に含まれる文書のベクトル要素の算術平均とした. すなわち, $\mathcal{K}_l$ の代表ベクトル $\boldsymbol{K}_l$ の j 番目の要素は

$$k_j^l \equiv \frac{1}{|\mathcal{K}_l|} \sum_{d_n \in \mathcal{K}_l} d_j^n \qquad (3)$$

3. 平均で大きな値を持っている語 t_j は, その文書群全体の特徴的な語とみなすことができる. よって次のように文書ベクトルを拡張する.

$$d_j'^n \equiv \max(d_j^n, k_j^0, k_j^1, ..., k_j^m) \qquad (4)$$

ただし、$k_j^0, \cdots, k_j^m$ は文書 d_n が属す文書群 $\mathcal{K}_0, \cdots, \mathcal{K}_m$ の代表ベクトルの第 j 要素である。以下, $d_j'^n$ によって構成される文書ベクトルを修正ベクトルと呼ぶ.

4.3 検索システムの概要

- 性能比較のために提案手法を含め, 4種類のシステムを用意した.
  - 単純な tf-idf
  - query expansion α
  - query expansion β
  - 提案手法
- 自然文で与えた query から形態素解析によって文書ベクトルと同様に体言・用言を抽出する.
- query の各語には重み付けが可能である. tf-idf と提案手法では全ての語の重みは1に固定, query expansion では後述の規則に従って重み付を行った.
- query ベクトルと文書ベクトルの内積によって文書の順位付けを行う.

4.3.1 query expansion α

- 2.1.1章「基本的な手法」の一例.
- シソーラスを用いて関連語を補う（実験では Express Finder のシソーラス辞書を用いた[7]）. 補われた語の重みは, 文書母集団における平均重要度とする.
- 元から query に存在する語の重みは, 全て α とする. $\alpha = 0.0005$ の時に最高性能が得られたので, これを本手法の代表結果とする.

4.3.2 query expansion β

- 2.1.2章「みなし正解中の重要語を用いる手法」の一例. SMART システム[1]の手法を参考にして, tf-idf による上位文書での重要語を補う.
- tf-idf における上位20文書に出現する語について重要度の平均を計算する.
- query に含まれている語以外で平均重要度の大きい25語を補う.
- 元から含まれている語, 補った語ともに20文書における平均重要度を新しい query ベクトルでの重みとする.
- 20文書 / 25語という値は幾つかの組合せから最高性能となるものを選んだ.

4.4 検索性能の評価方法

- 再現率・適合率を計算する.

$$(\text{再現率}) R_i = i/(\text{正解文書数}) \qquad (5)$$

$$(\text{適合率}) P_i = i/(\text{上位から } i \text{ 番目の正解文書の順位}) \qquad (6)$$

- 性能指標として11点平均適合率[8]を計算する.

■ 表 2　検索性能の比較 (積分適合率・11 点平均適合率)

太字は tf·idf に比べて高い値となったもの.
太字は提案手法よりも高い値となったもの.
検定の信頼度が (−) となっているのは tf·idf よりも低い性能であることを示す.

		正解集合 A				正解集合 A+B			
		tf·idf	提案手法	α	β	tf·idf	提案手法	α	β
P_{int}	query 1	.5994	.5575	**.6079**	.5102	.3249	**.3682**	**.3271**	**.3471**
	2	.6333	**.6857**	**.7829**	.5036	.3582	**.3932**	**.3992**	.3425
	3	.3245	**.3666**	.3153	.3154	.4311	**.5182**	**.4449**	**.4988**
	4	.0580	**.0615**	.0568	**.0746**	.3225	**.3842**	.3216	**.4216**
	5	.4321	**.5715**	**.4675**	.0494	.1531	**.1891**	.1502	.0908
11 点平均適合率の平均		.4358	**.4778**	**.4778**	.3096	.3394	**.3802**	**.3447**	**.3440**
tf·idf 比			**1.0964**	**1.0962**	.7105		**1.1202**	**1.0155**	**1.0134**
t 検定 信頼度			.9970	.9998	(−).9996		.9999	.6959	.1405
符号検定信頼度			.9355	1.0000	(−).9893		1.0000	.7744	.9355

ただし, $i_1 > i_2$ において $P_{i_1} < P_{i_2}$ となる場合は $P_{i_1} \leftarrow P_{i_2}$ という補間を行った.

- 手法間の性能の差異を統計的に検定する [8]. 手法は 11 点平均適合率の列において, 一方を tf·idf, 他方を提案手法あるいは query expansion α, β とした対標本に対し
  - ・対標本 t 検定 [9]
  - ・符号検定 [10]

  の 2 種の検定を行った.
- query ごとの特徴を計る目安として次の式による積分適合率を定義し, 各手法・各 query ごとに計算する.

$$P_{\mathrm{int}} \equiv \frac{1}{N} \sum R_i P_i \qquad (7)$$

4.5　結果

表 2 に実験の結果を示す. 順に

- 上 5 段は各 query の積分適合率（式 7）.
- 11 点平均適合率の 11 点の平均.
- tf·idf の平均 .4358 を 1 とした時の比
- 11 点平均適合率の 11 点を標本として tf·idf と対標本 t 検定を行ったときに帰無仮説「2 対が同じ母集団に属する」を棄却し得る最大信頼度.
- 11 点平均適合率の 11 点を標本として tf·idf と符号検定を行ったときに帰無仮説「2 対の母集団が等しい中央値を持つ」を棄却し得る最大信頼度.

である.

表 3 は手法の差が顕著に現れた query において, query expansion の各手法によって補われた語彙である. query 2 は「図形」という概念が多様な表現を持つことが原因という, 意味的曖昧性の典型的な例である. 手法αでは「図形」の関連語として「グラフィックス」「画像」などが補われており, これが効果的に働いたと考えられる. 一方βでは tf·idf での上位文書として, 非画像的なデータベース検索の話題が多かったために「テキスト」「鳴き声」などの語が補われ, 結果として適合率が下がってしまった. 手法βは, もともと正解文書が少ない場合や tf·idf での上位候補に正解が少ない場合にはみなし正解の精度が低くなるため, 不向きといえる.

提案手法では表 4 に示した文書のように, 「図形」や「パターン」という表現が『類似画検索』などのキーワードから補われ, 検索性能が向上している.

query 4, 5 は複数の話題を含んでいるために, 例えば query 4 で「ニューラルネット」あるいは「画像」のどちらかのみに関係した文書がノイズになっているという, 意味的曖昧性以外の問題も含んでいる. これに対しては, 2 つの話題の積集合に特徴的な語彙を query に補うことで大幅な性能向上が期待できる. 手法αでは文書母集団に対する重要度で固定的に重み付けを行ったために性能向上が見られなかった. 手法 β は query 5 では正解文書数の少なさ, tf·idf での適合率の低さから性能を落しているが, query 4 に対しては 2 つの話題を同時に含んだ文書から「エッジ」などの語彙を補うことに成功している. 提案手法は query を拡張しないので意味的曖昧性への効果のみとなっている.

■ 表 3　query expansion の具体例

query	α	β
2	グラフィックス, 画像, パタン, 型, 型紙, 原型, 図案, 文様, 柄, 模様, 紋様, 様式, サーチ, 探求, 探索, 調査	特徴, DB, 例示, コマ, 利用者, 画, TRADEMARK, 鳴き声, 文字列, 量, テキスト, 求める, 抽出, 動, 意味, 真似, 鳴く, メディア, 視覚, 構造, 主観的, 感じる, 対話, .
4	ニューラル, ネット, 運用, 活用, 使用, IP, グラフィックス, 画, 図形, 形質, 性格, 性質, 特性, 割合, 速度, 比, 率, 引き出し	認識, 感性, 領域, 分類, 学習, 類似, エッジ, 顔, デザイン, 文字認識, 方向, 色, RLC, 図形, 回転, しきい値, 求める, 基づく, VQ, 笑顔, 印象, 不変, 皮膚, テクスチャ
5	高速, フーリエ変換, コンカレント, パラレル, マルチ, 多重, 並行, プロセシング, プロセッシング, プロセッサ, プロセッサー, マルチプロセッサ, マルチプロセッシング, 処理	レイヤ, 高速, 処理, プロトコル, スレッド, プロセッサ, 高性能, 型, 評価, WS, メモリ, DSE, 共有, アービタ, 伝送路, DSP, ワークステーション, 分散, 市販, OS, 用いる, HXB, 通信, 照査

4.6　実験結果の考察

実験で確認された各手法の特徴をまとめると次のようになる.

- 手法αは 5 つの query の平均で約 2〜10%の適合率向上となった. 文書母集団とは独立した辞書を用いることで統計的手法以上の高再現率を達成し得る. 今回の実装のように語の重みが固定では検索者の意図する概念領域と一致しないこともある.
- 今回用いた query expansion は, 本来フィルタリングのように正解文書が与えられる場面で性能を発揮するものである. 正解数の少ないランク A に対しては性能が低下し, ランク B まで含めた場合で 1%程度の適合率向上となった. ただし, 意味的曖昧性への対策という以外に, 複数の話題を含む query に対して適切な表現を補う効果がある.
- 提案手法はランク A の正解集合で 10%, A+B で 12%程度適合率が向上した. これはランク A の文書, ランク B の文書, 不正解文書という順に得点が付けられているという望ましい状況を示している.

また, 3.2 章で述べた, query expansion よりもベクトル拡張の精度を高めるという効果が表 5 の例などに実際に表れており, 当初の目的に叶った結果だといえる. 統計的検定の結果から, 今後検索例を増やすことで tf·idf に対する提案手法の有意性をより明確に示せるものと考えている.

問題点を挙げると, 関連語ではあるが文書の内容を表してはいない語が補われることがある（表 6）. 提案手法のモデルを整理すると

$$d'^n_j \equiv f(d^n_j, k^0_j, k^1_j, ..., k^l_j) \qquad (4')$$
$$(\mathrm{d}_n \in \mathcal{K}_0) \cap ... \cap (\mathrm{d}_n \in \mathcal{K}_l)$$

となり, 文書群 $\mathcal{K}_l$, 代表ベクトル $\boldsymbol{K}_l$ の定義や関数 f を改善することで, ノイズとなる語を抑えることが検索性能的にも検索テーブルの膨張を抑える意味でも最大の課題である. 当面は

- 付与キーワードが完全一致の文書だけを関連性があるとしている点.
- キーワードの代表ベクトルの要素を算術平均としている点（式 3）.

この 2 点に注目しての改良を考えている. また, 統計的な現在の手法に辞書を用いた手法を組み合わせることも検討している.

実際の検索システムの利用状況ではシステムが出力した候補の様子を見ながら検索者が情報を追加入力するなどして検索を繰り返す relevance feedback が行われる. このような状況での性能評価も必要だろう.

5　おわりに

本論文では自然言語の意味曖昧性が情報検索において問題となっていることを取り上げ, 文書関連性を用いて文書ベクトルを拡張することで検索性能を向上させる手法を提案した.

今回の性能評価実験で用いた正解集合ではランク A の文書数が少なく, query 自体の数も 5 件で, 手法の特徴を詳細に検討するのに十分な規模であるとはいえない. 現在, 学術文献を対象とした約 30 万件という大規模な IR 用日本語テストコレク

■ 表 4　提案手法で他手法より上位に順位付けられた正解文書の例

query	タイトル	キーワード	順位 tf·idf →提案手法
4	階層透過型特徴抽出によるパターン認識	ニューラルネットワーク, パターン認識, 視覚情報処理, 特徴抽出	128 → **35**
5	並列ベクトル計算機「数値風洞」による FFT プログラムの性能評価	高速フーリエ変換, 性能評価, 並列ベクトル計算機, 計算流体力学, 数値風洞	120 → **13**

■ 表 5　提案手法で他手法より下位に順位付けられた不正解文書の例

query	タイトル	キーワード	順位　tf·idf → 提案手法, α, β
4	プロセスの特徴を考慮した動的負荷分散についての一考察	分散オペレーティングシステム, 動的負荷分散	91 → **164**, 250, 57
5	可変構造型並列計算機の並列オペレーティング・システム	マルチプロセッサ, スレッド, スケジューリング, オペレーティング・システム, アドレス変換, スタック	158 → **255**, 9, 103

■ 表 6　提案手法で他手法より上位に順位付けられた不正解文書の例

query	タイトル	キーワード	順位 tf·idf →提案手法
1	国際 LAN 間接続を想定した TCP スループット特性の評価	TCP, スループット, LAN 間接続	930 → **26**
4	ニューラルネットを利用した文字認識実験の一検討	階層型ニューラルネットワーク, 文字認識, ベクトル量子化法, 特徴抽出	195 → **18**

ションの整備が学術情報センターにおいて進められており [12], テスト版 30 件の query を利用しての評価実験を計画している.

参考文献

[1] Chris Buckley, Amit Singhal, Mandar Mitra, "Using Query Zoning and Correlation Within SMART : TREC 5," 1996.

[2] Hyun-Kyu Kang, Key-Sun Choi, "Two-level document ranking using mutual information in natural language information retrieval," *Information Processing & Management*, Vol.33, No.3, pp.289–306, 1997.

[3] Deerwester,S., Dumais,S.T., Landauer,T.K., Furnas,G.W. and Harshman,R.A. "Indexing by Latent Semantic Analysis," *J. of the American Society for Information Science*, Vol.41, No.6, pp.391–407, 1990.

[4] 宮原 隆行, 清木 康, 北川高嗣 "意味の数学モデルによる意味的連想検索の高速化アルゴリズムとその実現方式," 情処論, Vol.38, No.7, pp.1399 – 1411, July 1997.

[5] 日本語形態素解析器 茶筌 (ChaSen), `http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html`

[6] 松村 敦, 高須 淳宏, 安達淳, "構造化インデクスを用いた情報検索手法の評価," 信学技法 DE98-2, pp.7 – 14, May 1998.

[7] NTT アドバンストテクノロジ Express Finder / シソーラス辞書 '97 年版.

[8] "Introduction to Modern Information Retrieval," Salton,G. and McGill,M., *McGraw-Hill*, 1983

[9] "統計," 共立講座 21 世紀の数学 14, 竹村 彰通, 共立出版, 1997.

[10] "工業統計学," 村上 征勝, 朝倉書店, 1985.

[11] Ellen M. Voorhees, Donna Harman, "Overview of the Fifth Text REtrival Conference (TREC-5)," 1996.

[12] "NTCIR : NACSIS Test Collection Project [Poster]," Kando,N., Koyama,T., Oyama,K., Kageura,K., Yoshioka,M., Nozue,T., Matsumura,A. and Kuriyama,K., *the 20th Annual Collquium of BCS-IRSG*, March 25-27, 1997, Autrans, France.