社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－DBS－134　(II)　(68)
2004／7／14

意味的連想検索方式における意味表現ベクトルを対象とした学習機構の実現式

大橋英博[†]　清木康[††]

本稿では意味的連想検索方式において用いられる意味表現ベクトルを対象とし, 検索精度を改善するための学習機構の実現を示す. 本方式は意味表現ベクトルの要素反転により重要なメタデータを特定することにより, 意味表現ベクトルを学習させる方式である. 本方式をデータベースを対象とした意味的連想検索方式に適用することにより, 検索者の要求に応じた検索結果の学習が可能となる.

An Implementation Method of Learning Mechanisms for Meaning Expression Vectors on A Semantic Associative Search Mechanism

Hidehiro Ohashi[†] and Yasushi Kiyoki[††]

In this paper, we present an implementation method of learning mechanisms for meaning expression vectors in semantic associative search. This method is used to create appropriate meaning expression vectors by checking and modifying each element of a vector as a learning process. By using modified meaning expression vectors, we can realize semantic associative search with high quality. Our method is used to improve the precision for the retrieval result in semantic associative search for databases.

1. はじめに

現在, ネットワーク上にはアクセス可能なデータベースが数多く存在している. 大量の情報源から有用な情報にアクセスするために, データベース検索システムが多くの場面で利用されている. データベース検索システムには, 検索者の検索要求を満たす検索結果を返すことが求められる. 一般に, 検索者の検索要求を完全に満たすデータベース検索システムを最初から構築することは困難である. 検索者の検索要求を満たす検索精度の高いデータベース検索システムを実現するためには, システム構築後に, データベース検索システムを検索者の要求に合わせてを学習させる機構が必要である. データベース検索システムにおける学習とは検索者の要求にに検索結果が一致するようにデータベースを修正することである.

† 慶應義塾大学大学院政策・メディア研究科
Graduate School of Media and Governance, Keio University
e-mail：hohashi@sfc.keio.ac.jp
†† 慶應義塾大学環境情報学部
Faculty of Environmental Information, Keio University
e-mail：kiyoki@sfc.keio.ac.jp

ベクトル空間モデルが, 情報検索方式の一つとして提案されている. ベクトル空間モデルは, 検索の対象となる情報と検索者の検索質問を表す情報とをそれぞれベクトルとして表現する. ベクトル空間モデルでは, これらのベクトルの相関の高さによって, 検索質問と検索の対象となる情報との近さを計量する. ベクトル空間モデルにおける学習には, 2種類の方法が考えられる. 第 1 の学習方法は, 検索質問のベクトルを検索対象のベクトルに近づける方法である. 第 2 の学習方法は検索対象のベクトルを検索質問のベクトルに近づける方法である. これらの方法は学習の結果が及ぼす影響範囲に違いがある. 第 1 の学習方法では検索質問のベクトルを修正するために, 当該検索質問について全ての検索対象のベクトルの相関量が変わる. この方法による影響は, 当該検索質問を使用した検索に限られる. 第 2 の学習方法では検索対象のベクトルを修正するため, 当該検索対象のベクトルの相関量だけが変わる. 第 2 の学習方法では, 他の検索質問で検索を行った場合においても, 当該検索対象ベクトルの相関量が変わることになる.

2. 意味的連想検索方式の概要

ベクトル空間モデルによる検索方式の一つとして意味的連想検索方式[1)~4)]が提案されている．この方式は検索者の文脈を解釈する検索方式である．この方式における文脈とは，検索語のもつ多義性を解消するために検索語と共に与えるものであり, 検索語の意味を確定するための情報である. この文脈は以降で説明する特徴語群によって特徴付けされたベクトル（文脈ベクトル）として表現される．同じく検索対象の情報も特徴語群によって特徴付けたベクトル（検索対象メタデータベクトル）として表現する．

意味的連想検索方式の特徴は, 専門家の知識を意味空間として体系化し, 検索者が入力した検索質問と検索対象の情報との意味的な類似性を計量する空間として使用する点にある. ここで, 専門家の知識（専門知識）とは, 意味空間を形成するための特徴付きベクトルの作成に用いる専門分野辞書および事典に記述されている当該分野に関する情報をさす. 意味的連想検索方式により, 検索者の文脈を専門家の知識体系に基づいて解釈し, 検索質問と検索対象の情報との意味的な類似性に基づいて, 情報検索を行うことが可能となる.

意味的連想検索方式は専門知識を意味空間として体系化する. ここで意味空間とは, 意味の数学モデル[1)~4)]に基づいて構成されたn次元正規直交ベクトル空間をさす. 意味の数学モデルでは, 当該分野の専門用語と, 専門用語を説明するための単語（特徴語）を使用して専門用語の特徴付きベクトルを作成する. 特徴付きベクトルは各専門用語ごとに作成する. 特徴付きベクトル群から意味空間生成用マトリクスを作成する．意味空間生成用マトリクスを直交化し, n次元正規直交ベクトル空間（意味空間）を作成する. 意味的連想検索方式では, 文脈ベクトルによって意味空間上の部分空間を選択する．この部分空間上に検索対象メタデータベクトルに写像する．部分空間上の検索対象メタデータベクトルのノルムにより, 文脈ベクトルと検索対象メタデータベクトルとの意味的な類似性を計量する．

このように意味的連想検索方式においては, 文脈ベクトルによって選択された部分空間上に検索対象メタデータベクトルを写像してノルムを計量する．検索対象メタデータベクトルに対する学習は部分空間に写像する前の検索対象メタデータベクトルに対して行う．検索者の要求に一致する検索結果を得るためには, 部分空間上で適切なノルムを得るように, 部分空間に写像する前の検索対象メタデータベクトルを修正する機構が必要となる．

3. 本研究の目的

本研究の目的は意味的連想検索方式の検索結果を検索者の要求に応じて学習させる機構を実現することである．この目的を達成するために, 意味表現ベクトルを意味的連想検索方式の特徴に応じて効率的に修正する機構を実現する．

3.1 意味表現ベクトル

意味的連想検索方式では, 当該分野の専門用語と, 専門用語を説明するための単語（特徴語）を使用して専門用語の特徴付きベクトルを作成する．この特徴付きベクトルを用いて, 意味空間生成のためのマトリクスを構成する．また, 文脈ベクトルや検索対象メタデータベクトルも特徴付きベクトルによって表現される．これら文脈ベクトルや検索対象メタデータベクトルを総称して意味表現ベクトルとよぶ．本研究では, この意味表現ベクトルを対象とした学習機構を実現する．

3.2 本研究における学習の対象

本研究では意味表現ベクトルの中で, 検索対象メタデータベクトルを学習の対象とする．検索対象メタデータベクトルを学習の対象とすることで, 学習の影響を当該検索対象メタデータベクトルだけに限定することができる．

3.3 意味的連想検索方式

3.3.1 意味の数学モデルの基本構成

意味の数学モデルによる意味的連想検索方式[1)~4)]のメディアデータ検索方式の概要を示す．

(1) **メタデータ空間 $\mathcal{MDS}$ の設定**

検索対象となるメディアデータをベクトルで表現したデータをマッピングするための正規直交空間 (以下，メタデータ空間 $\mathcal{MDS}$) を設定する．

(2) **メディアデータのメタデータをメタデータ空間 $\mathcal{MDS}$ へ写像**

設定されたメタデータ空間 $\mathcal{MDS}$ へメディアデータのメタデータをベクトル化し写像する．これにより，同じ空間に検索対象データのメタデータがメタデータ空間上に配置されることになり，検索対象データ間の意味的な関係を空間上での距離として計算することが可能となる．

メディアデータ P には，メタデータとして t 個の印象語 $\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_t$ が以下のように付与されていることを前提としている．

$$P := \{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_t\}. \tag{1}$$

各印象語 $\mathbf{w}_i$ は，ベクトル表現された特徴 $f_{i1}, f_{i2}, \cdots, f_{in}$ を持っている．

$$\mathbf{w}_i := (f_{i1}, f_{i2}, \cdots, f_{in}). \quad (2)$$

各メディアデータは，メタデータとして付与されている t 個の印象語が合成されベクトル表現された後，メタデータ空間 $\mathcal{MDS}$ へ写像される．

(3) **メタデータ空間 $\mathcal{MDS}$ の部分空間 (意味空間) の選択**

検索者は与える文脈を複数の単語を用いて表現する．検索者が与える単語の集合をコンテキストと呼ぶ．このコンテキストを用いてメタデータ空間 $\mathcal{MDS}$ に各コンテキストに対応するベクトルを写像する．これらのベクトルは，メタデータ空間 $\mathcal{MDS}$ において合成され，意味重心を表すベクトルが生成される．意味重心から各軸への射影値を相関とし，閾値を超えた相関値 (以下，重み) を持つ軸からなる部分空間 (以下，意味空間) が選択される．

(4) **メタデータ空間 $\mathcal{MDS}$ の部分空間 (意味空間) における相関の定量化**

選択されたメタデータ空間 $\mathcal{MDS}$ の部分空間 (意味空間) において，メディアデータベクトルのノルムを検索語列との相関として計量する．これにより，与えられたコンテキストと各メディアデータとの相関の強さを定量化している．この意味空間における検索結果は，各メディアデータを相関の強さについてソートしたリストとして与えられる．

また，メディアデータを特徴づける特徴の数が多い場合，どのような意味空間が選ばれても，意味空間におけるメディアデータのノルムが大きくなる傾向がある．そのため，本来，文脈との相関が強いと考えられるメディアデータベクトルのノルムよりも，特徴の数が多いメディアデータベクトルのノルムが大きくなってしまい，適切な抽出が行われないことがある．そのため，メタデータ空間でのメディアデータベクトルを 2 ノルムで正規化している．

4. 検索対象メタデータベクトルを対象とした学習方式

4.1 本学習機構の概要

本学習機構は以下の手順で学習を行う．

(1) 検索対象メタデータベクトルの各 1 要素の反転によるノルムの変化の分析

(2) 重要な特徴語の追加および削除

次に各手順について説明する．

4.2 検索対象メタデータベクトルの各 1 要素の反転によるノルムの変化の分析

検索対象メタデータベクトルは，特徴語によって特徴付けされている．検索対象メタデータベクトルの各要素は，各特徴語に対応している．検索対象メタデータベクトル T が示す情報が特徴語 f_i によって特徴付けられる場合，検索対象メタデータベクトル T の第 i 要素は 1 となる．そうでない場合，第 i 要素は 0 となる．1 要素ずつ反転させたときの検索対象メタデータベクトル T に対するノルムを調べることにより，ノルムを大きくする特徴語およびノルムを小さくする特徴語を見つけることができる．1 要素の反転操作は，検索対象のメタデータについて，対応する特徴語を追加もしくは削除することに相当する．検索対象メタデータベクトルの第 i 要素を 0 から 1 に反転させることは検索対象のメタデータに特徴語 f_i を追加することに相当する．検索対象メタデータベクトルの第 i 要素を 1 から 0 に反転させることは，検索対象のメタデータから特徴語 f_i を削除することに相当する．図 1 で例を示す．図 1 では現在の検索対象メタデータベクトル T と，T を部分空間に写像したベクトル T_m のノルム N_{T_m}，および N_{T_m} の検索結果中における順位 R_{T_m} を示している．例えば，検索対象メタデータベクトルが 3 語の特徴語によって特徴付けされているとする．この場合，検索対象メタデータベクトルは 3 次元となる．現在の検索対象メタデータベクトル T を T = (1 0 1) とする．このとき，各特徴語に対応する要素を反転させたときの，部分空間における検索対象メタデータベクトル T_m のノルム N_{T_m} を変化を調べる．元のベクトル T の各要素を 1 要素ずつ反転させると，新しく 3 ベクトル $T^{(1)}$〜$T^{(3)}$ を作ることができる．この 3 ベクトル $T^{(1)}$〜$T^{(3)}$ の部分空間のノルム $N_{T_{m_1}}$〜$N_{T_{m_3}}$ および順位 $R_{T_{m_1}}$〜$R_{T_{m_3}}$ を調べる．元のベクトル T のノルム N_{T_m} と 1 要素ずつ反転させたベクトル $T^{(i)}$ のノルム $N_{T_{m_i}}$ を比較する．$N_{T_{m_i}} > N_{T_m}$ となるとき，特徴語 f_i を反転させることにより，検索対象メタデータベクトルのノルムが大きくなることが分かる（図 1 のグループ 1）．同様に $N_{T_{m_i}} < N_{T_m}$ となるとき，特徴語 f_i を反転させることにより，検索対象メタデータベクトルのノルムが小さくなることが分かる（図 1 のグループ 2）．$N_{T_{m_i}} = N_{T_m}$ となるとき，特徴語 f_i を反転させても，検索対象メタデータベクトルのノルムが変化しないことが分かる（図 1 のグループ 3）．このように，元の検索対象メタデータベクトル T の各要素を反転させることにより，特徴語

ごとに検索対象メタデータベクトル T_m のノルムに対する影響を調べることができる．この分析を行うことにより，検索対象メタデータベクトルのノルムを大きくする特徴語群（グループ 1）や，ノルムを小さくする特徴語群（グループ 2）に分類できる．各要素を反転した後のノルムの値を，現在の検索結果と照らし合わせることにより，各要素を反転した後の当該検索対象メタデータベクトルの順位を知ることができる．

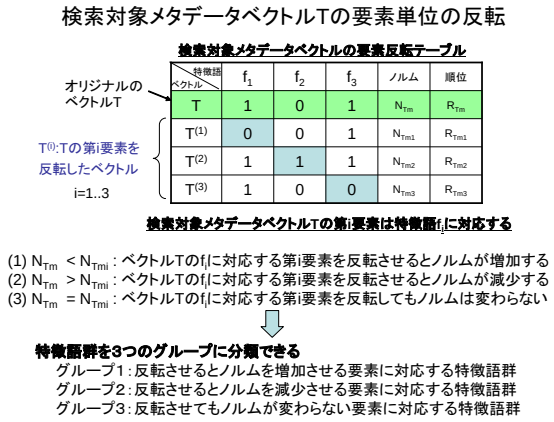

図 1　検索対象メタデータベクトルの各要素の反転とノルムの変化

4.3　重要な特徴語の追加や削除

検索対象メタデータベクトルの各要素の反転を行うことにより，ノルムを大きくする特徴語群と，ノルムを小さくする特徴語群に分けることができる（反転してもノルムが変動しない特徴語群は学習には影響がないため，ここでは無視する）．また，各要素を反転した後の順位も同時に把握することができる．現在の検索対象メタデータベクトルの順位 R_{T_m} を上げたい場合は，$R_{T_m} < R_{T_{m_i}}$ となる第 i 要素を選び，第 i 要素を反転させる．第 i 要素を反転させることで，第 i 要素に対応する特徴語 f_i が検索対象のメタデータから追加もしくは削除される．当該検索対象メタデータベクトルの順位が求める順位と一致するまで，要素の反転をを繰り返す．検索対象メタデータベクトルの順位が求める順位と一致した時点で，要素の反転を終了する．これにより，検索対象メタデータベクトルを検索者の要求に一致させることが可能となる．

5．実　　験

本研究では本学習方式による学習実験を行った．ここでは学習実験の方法と結果について述べる．

5.1　実 験 方 法

医療に関する意味空間（医療意味空間）を用いた学習実験を行った．実験では，あらかじめ用意した検索質問と正解ドキュメントの組を使用して，再現率と適合率を計算した．その後，本方式によって学習を行った．学習した結果について，再度再現率と適合率を計算した．学習前に比べて学習後の再現率と適合率が大きく改善されることを確認した．実験では意味的連想検索方式におけるノルム系検索を使用した．適合率と再現率は図 2 の計算式で算出した．

$$\text{再現率}(\%) = \frac{\text{検索結果として得られた中の正解数}}{\text{全正解数}} \times 100$$

$$\text{適合率}(\%) = \frac{\text{検索結果として得られた中の正解数}}{\text{検索結果数}} \times 100$$

図 2　再現率・適合率の計算式

5.2　実験に使用したデータ

医療意味空間を，医療に関する専門辞書を用いて構築した．医療意味空間は 436 次元の正規直交ベクトル空間である．検索対象として 201 件の医療に関するドキュメントを使用した．

学習実験を行うための正解セットとして，10 語の検索質問を用意した．また各検索質問に対して 5 件の正解ドキュメントを用意した．以下に実験で使用した検索質問を示す．本実験では，各検索質問に対する検索結果の上位 5 件を用いて，再現率および適合率を算出した．

表 1　実験で使用した検索質問

番号	検索質問
1	エイズ
2	遺伝子
3	胃潰瘍
4	がん
5	アルツハイマー病
6	パーキンソン病
7	花粉症
8	高血圧
9	肝臓障害
10	呼吸困難

5.3　実 験 結 果

学習前の平均再現率・平均適合率と，学習後の平均

再現率・平均適合率を表 2, 3 に示す. 平均再現率・平均適合率は 10 語の検索質問における再現率・適合率の算術平均によって算出した. 学習前の平均再現率は 72(%), 平均適合率は 72(%) であった. 学習の結果, 平均再現率は 92(%), 平均適合率は 92(%) に上昇した.

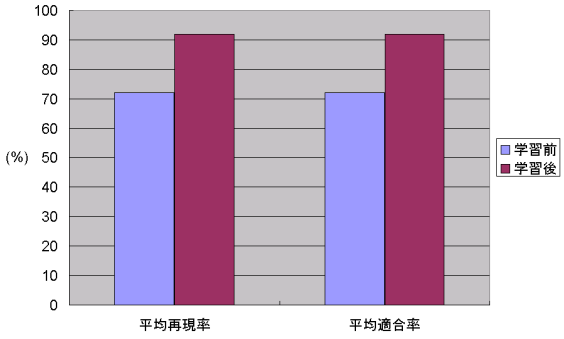

図 3 学習による平均再現率と平均適合率の変化

表 2 学習による再現率の変化

番号	検索質問	学習前再現率 (%)	学習後再現率 (%)
1	エイズ	80	100
2	遺伝子	80	100
3	胃潰瘍	80	100
4	がん	80	100
5	アルツハイマー病	60	80
6	パーキンソン病	60	80
7	花粉症	40	100
8	高血圧	80	80
9	肝臓障害	80	100
10	呼吸困難	80	80
	平均再現率	72	92

表 3 学習による適合率の変化

番号	検索質問	学習前適合率 (%)	学習後適合率 (%)
1	エイズ	80	100
2	遺伝子	80	100
3	胃潰瘍	80	100
4	がん	80	100
5	アルツハイマー病	60	80
6	パーキンソン病	60	80
7	花粉症	40	100
8	高血圧	80	80
9	肝臓障害	80	100
10	呼吸困難	80	80
	平均適合率	72	92

6. 考　　察

本実験では, 10 語の検索質問中, 6 語について, 再現率と適合率を大きく (100(%) に近く) 改善にすることができた（表 2, 表 3）. 学習後の平均再現率, 平均適合率も 92(%) となっている. これらのことから, 本方式によって, 効率的な学習が行えることを確認できた.

表 4 に示す 4 語の検索質問による検索については, 学習を行った後においても, 再現率と適合率が 100(%) にはならなかった. これは, ある検索質問における学習が別の検索質問の学習に対して影響を及ぼしたためである. 実験では, 検索質問「アルツハイマー病」によって得られた検索結果の中で, ドキュメント「98032032」の学習前の順位は 8 位となっていた. このドキュメントは本来 5 位以内にランキングされるべきドキュメントであったため, 本方式により, 5 位となるように学習させた. しかし, 検索質問「パーキンソン病」に対する学習を行った後で, もう一度「アルツハイマー病」で検索を行ってみると, ドキュメント「98032032」は 6 位にランキングされた. 5 位にはドキュメント「980404240」がランキングされた. 本来 5 位となるべきドキュメント「98032032」が 6 位にランキングされたため, 再現率と適合率が 100(%) にならなかった.

原因は, 次のとおりである. 「アルツハイマー病」による学習を行った時点では, ドキュメント「98032032」は 5 位に正しくランキングされていた. しかし, 「パーキンソン病」の学習を行った際, ドキュメント「980404240」に対して特徴語の追加を行ったため, 「アルツハイマー病」による検索においてもドキュメント「980404240」の検索対象メタデータベクトルのノルムが大きくなった. このため, 本来 5 位にランキングされるべきドキュメント「98032032」のノルムよりも, ドキュメント「980404240」のノルムが大きくなった. この結果, 「アルツハイマー病」における検索においてドキュメント「980404240」が 5 位となり, ドキュメント「98032032」が 6 位となった.

このように, 本学習方式は検索対象メタデータベクトルを学習の対象としているため, ある検索質問に対する学習の影響が別の検索質問に及ぶ可能性がある. この場合, 検索対象メタデータベクトルだけでなく文脈ベクトルも対象とした学習を行う必要がある.

7. 結　　論

本稿では, 意味的連想検索方式の特徴に応じて, 検索対象メタデータベクトルを効率的に修正する機構を実現した. 検索対象メタデータベクトルの各要素の反

表 4 再現率・適合率が 100(%) にならなかった検索質問

番号	検索質問	学習後再現率 (%)	学習後適合率 (%)
5	アルツハイマー病	80	80
6	パーキンソン病	80	80
8	高血圧	80	80
10	呼吸困難	80	80

転を行うことにより, 特徴語を追加または削除した時の, 検索対象メタデータベクトルのノルムの変化を調べることができる. この情報をもとにした特徴語の追加または削除を行うことにより, 検索者の要求に一致するように, 検索結果を学習させることが可能となる. 本方式により, 意味的連想検索方式の検索結果を検索者の要求に応じて学習させることが可能となった.

参 考 文 献

1) Kitagawa, T. and Kiyoki, Y.: The mathematical model of meaning and its application to multidatabase systems, Proceedings of 3rd IEEE International Workshop on Research Issues on Data Engineering: Interoperability in Multidatabase Systems, pp.130-135, April 1993.

2) Kiyoki, Y., Kitagawa, T. and Hayama, T.: A metadatabase system for semantic image search by a mathematical model of meaning, ACM SIGMOD Record, Vol. 23, No. 4, pp.34-41, 1994.

3) Kiyoki, Y.,Kitagawa, T. and Hitomi, Y.: A fundamental framework for realizing semantic interoperability in a multidatabase environment, Journal of Integrated Computer-Aided Engineering, Vol.2, No.1, pp.3-20, John Wiley & Sons, Jan. 1995.

4) 清木 康, 金子 昌史, 北川 高嗣: 意味の数学モデルによる画像データベース探索方式とその学習機構, 電子情報通信学会論文誌,D-II,Vol.J79-D-II,No.4,pp.509-519, 1996.