2005－DBS－137（Ⅰ）（3）
2005／7／13

書籍の索引部を用いた検索空間生成方式における空間統合方式

中西崇文[†]　岸本貞弥[†]
櫻井鉄也[††]　北川高嗣[††]

本稿では，様々な情報源から抽出された語と語の関係を表す各検索空間を対象とした検索空間統合方式について述べる．特定分野を対象としたベクトル空間モデルを用いた検索機構を実現するためには，その分野を対象とした語と語の関係を計量可能な検索空間を生成する必要がある．我々はこれまで書籍の索引を用いた検索空間生成方式を提案してきた．１つの書籍の索引では，書籍の性質から，検索対象となるメタデータ空間の扱う語彙数が少なくなる傾向にある．一方，これまで様々な情報源からその関係を抽出し，検索空間を生成する方式が提案されてきた．これらの空間を統合することにより，その問題が解決される．本方式を含む書籍の索引部によるメタデータ空間生成方式は，学術分野だけでなく，趣味など，幅広い分野における，メディアデータ検索，ドキュメント検索に応用できると考えられる．

An Integration and Extension Method of a Retrival Space utilizing an Indexe of a book

Takafumi NAKANISHI,[†] Sadaya KISHIMOTO,[†]
Tetsuya SAKURAI[††] and Takashi KITAGAWA [††]

In this paper, we present an integration method of retrival spaces. This method make it possible to integrate retrival spaces based on the relation between words. In order to realize search method utilizing a vector space model for a specific field, it is necessary to construct a retrival space for the field. The number of the vocabularies, which the retrival space constructed from a index of a document can express is restricted. The problem is solved by this method that uses other retrival spaces. It is thought that the retrival space constructed method by the index part of the documents containing this method is applicable to mediadata and document search for broad fields, such as a field of not only a scientific field but a hobby.

1. はじめに

コンピュータネットワーク上に特定分野を対象とした多種多様な情報群が広域に遍在しつつある．これらの情報を対象とした，情報獲得効率の低さが大きな問題となっている．そのため，特定分野における情報群を対象とした，高度な検索方式が重要となっている．

これまで，広域に偏在する情報群を対象とした高度な検索方式として，ベクトル空間モデル [1] による検索方式が提案されている．広義のベクトル空間モデル(以下，ベクトル空間モデル) とは，検索対象，および問い合わせを一定の特徴を用いてベクトル化し，それらを内積などの計量を行うことにより類似度を求める検索方式である．ここで，本定義には，LSI(Latent Semantic Indexing)[2),3)]，および我々が提案してきた検索方式である意味の数学モデル [4),5),7)] も含む．

ベクトル空間モデルを用いて各特定分野の質の高い情報を検索するためには，その特定分野に用いられる特徴となりうる専門用語同士の関係を抽出し，専門用語同士の計量が可能な検索空間を構築する必要がある．

これまで語と語の関係を計量するための検索空間の生成方式として，検索対象のメタデータを検索空間生成のためのメタデータとして用いる方法 [2),3)]，Longman Dictionary of Contemporary English(以下，Longman)[6)] という英英辞典や用語辞典を用いる方法 [5),8),9)] が提案されている．勿論，上記のように自動的，半自動的に実現する方式だけでなく，専門家などの人手で構築する方式も考えられる．

† 筑波大学大学院システム情報工学研究科，つくば市
Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan
e-mail：takafumi,kishimoto@mma.cs.tsukuba.ac.jp
†† 筑波大学大学院システム情報工学研究科，つくば市
Faculty of Systems and Information Engineering, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan
e-mail：sakurai,takashi@cs.tsukuba.ac.jp

また，我々は，これまで，特定分野の書籍の索引部を用いて語の関連を計量する専門分野を対象とした検索空間を生成する方式 [10] について研究を進めてきた．本方式は，検索空間を生成したい対象となる特定分野のことについて書かれた書籍を準備し，索引を参照することで，その特定分野の検索空間を容易に少ない手間で生成が可能となる．しかしながら，一般的に書籍 1 冊の索引に収録されている語数は高々数百から数千であり，空間上で用いることができる語彙数が少なくなってしまう問題点があった．

書籍の索引から抽出された語と語の関係を表す検索空間と他の異種の情報源から抽出された検索空間生成を統合し空間生成を行うことにより，書籍の索引の語彙数の少なさを解消し，かつ索引を用いることから，容易に，専門知識を必要せず，少ないコストで，メタデータ空間の生成ができる．

本稿では，様々な情報源から抽出された語と語の関係を表す各検索空間を対象とした検索空間統合方式について示す．特に，本稿では，書籍の索引部を用いて生成した検索空間を中心に考察する．

2. ベクトル空間モデルの概要

本節では，ベクトル空間モデルについて示す．なお，ここで示すのは広義のベクトル空間モデルであり，LSI や意味の数学モデルを考慮に入れ，なるべく一般化した形で記述を試みる．

2.1 ベクトル空間モデルの基本構成

検索対象の特徴を表す検索対象ベクトル $\mathbf{d}_i$ は，特徴づけとして用いる語に付与する重み $d_{i1}, d_{i2}, \cdots, d_{in}$ を要素として構成される n 次元ベクトルで表される．

$$\mathbf{d}_i = (d_{i1}, d_{i2}, \cdots, d_{in}) \tag{1}$$

一方，問い合わせを表す問い合わせベクトル $\mathbf{o}$ は，検索対象ベクトルで特徴づけとして用いた語を用いて特徴づけを行うことによって n 次元ベクトルで表される．

$$\mathbf{o} = (o_1, o_2, \cdots, o_n) \tag{2}$$

これらに対して，計量系 $\rho(\mathbf{d}_i; \mathbf{o})$ を定義し，問い合わせと各検索対象の類似度を導出する．

ここで，このままで計量を行うと，検索対象ベクトル，および問い合わせベクトルの特徴づけとして用いた語同士の関連を取り入れていない状態であるため，類似の計量がうまく行われない．

例えば，検索対象データ 1 のメタデータとして「本」「入門」，検索対象データ 2 のメタデータとして「書籍」「入門」と特徴づけられている．そのときに「書籍」という問い合わせを発行したとする．これらについて，検索対象データ 1 の検索対象ベクトルを $\mathbf{d}_1 = (1, 1, 0)$，検索対象データ 2 の検索対象ベクトルを $\mathbf{d}_2 = (0, 1, 1)$，問い合わせベクトルを $\mathbf{o} = (0, 0, 1)$ と表すことが出来る．ここで，内積で計量するとすると，検索対象データ 1 の類似度が 0 となる．実際は「書籍」と「本」は意味的に関連が大きいはずであるが，そのような情報が与えられていないために，相関を見出せない．

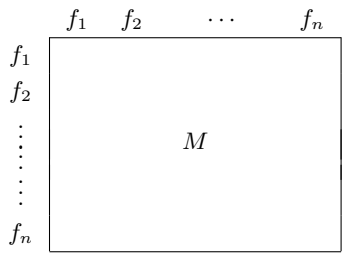

図 1 データ行列 M によるメタデータの表現.
Fig. 1 Metadata represented in data matrix M.

ここで，特徴として用いる語同士の関連の関連を導入する．検索対象ベクトル，および問い合わせベクトルの特徴として用いた n 個の語 $f_1, f_2, \cdots, f_n$ の各関係を示す図 1 のような n 行 n 列のデータ行列 M を用意する．この行列は実対称行列となる．

実対称行列 M を固有値分解する．

$$M^T M = Q \begin{pmatrix} \lambda_1 & & & & \\ & \ddots & & & \\ & & \lambda_\nu & & \\ & & & 0 & \\ & & & & \ddots 0 \end{pmatrix} Q^T,$$

$0 \leq \nu \leq n.$

ここで行列 Q は，

$$Q = (\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_n)$$

である．この $\mathbf{q}_i (i = 1, \cdots, n)$ は，相関行列の正規化された固有ベクトルである．この固有値は全て実数であり，その固有ベクトルは互いに直交している．

以上で導出された非ゼロ固有値に対応する固有ベクトルによって形成される正規直交空間が検索空間 $\mathcal{MDS}$ となる．この空間は，ν 次元ユークリッド空間となる．

$$\mathcal{MDS} := span(\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_\nu).$$

$\{\mathbf{q}_1, \cdots, \mathbf{q}_\nu\}$ は $\mathcal{MDS}$ の正規直交基底である．

この検索空間 $\mathcal{MDS}$ に検索対象ベクトル，問い合わせベクトルを写像するため，検索空間 $\mathcal{MDS}$ 内でフーリエ展開し，フーリエ係数を求める．つまり，検索空間の基底である各非ゼロ固有ベクトル $\mathbf{q}_i (i = 1, \cdots, \nu)$ との内積を求める．

検索対象ベクトル $\mathbf{d}_i$ と各非ゼロ固有ベクトル $\mathbf{q}_i (i =$

$1, \cdots, \nu$) との内積を求めた結果を要素とした，検索空間 $\mathcal{MDS}$ 内に写像された検索対象ベクトル $\widehat{\mathbf{d}}_i$ を以下に表す．

$$\widehat{\mathbf{d}}_i := (d'_{i1}, d'_{i2}, \cdots, d'_{i\nu}). \quad (3)$$

同様に問い合わせベクトル $\mathbf{o}$ と各非ゼロ固有ベクトル $\mathbf{q}_i (i = 1, \cdots, \nu)$ との内積を求めた結果を要素とした，検索空間 $\mathcal{MDS}$ 内に写像された検索対象ベクトル $\widehat{\mathbf{o}}_i$ を以下に表す．

$$\widehat{\mathbf{o}} := (o'_1, o'_2, \cdots, o'_{i\nu}). \quad (4)$$

検索空間 $\mathcal{MDS}$ 内に写像された検索対象ベクトル，問い合わせベクトルで，類似度 $\rho(\widehat{\mathbf{d}}_i; \widehat{\mathbf{o}})$ を導出する．

ここで，語と語の関連をデータ行列として，n 行 n 列の実対称行列 M で表すことを前提としていたが，データ行列は m 行 n 列の実行列 (つまり，ある物事を特徴として用いられている特徴づけに用いられる語で特徴づけしている行列) であっても，n 行 m 列の実行列 (つまり，特徴づけに用いられる語が他の特徴量で特徴づけられている行列) であっても構わない．

例えば，m 行 n 列の実行列を A と置くと，行列 A の相関行列 A^TA は n 行 n 列の語と語の関係を表す実対称行列となり，上記の操作を行うことができる．

なお，相関行列 A^TA の固有ベクトルは，行列 A の右特異ベクトルと等価であり，相関行列 AA^T の固有ベクトルは，行列 A の左特異ベクトルと等価である．つまり，LSI における特異値分解は，語同士の関係を計量可能な検索空間を張るための基底を求めている操作である．

これらにより，ベクトル空間モデルにおいて，語と語の関係であるデータ行列，つまり語同士の関係を表すメタデータを何処から抽出するかが重要となる．

2.2 ベクトル空間モデルを実現するためのメタデータ

前節の内容を整理すると，ベクトル空間モデルで必要となるメタデータの種類として以下の 3 つが挙げられる．

(1) 検索空間生成のためのメタデータ

語同士の関係を計量可能な検索空間を生成するためのメタデータであり，検索空間を構成するためのデータ行列となりうるデータである．検索対象や問い合わせのメタデータで特徴づけに用いられる各語同士の関連を記述する．このデータは，前節で示すように，直接各語と語の関係を記述なくてよく，ある概念や事柄を介して関係を記述しても構わない．介した概念や事柄によって，語同士のどういう関係を計量する検索空間になるかが決まる．

(2) 検索対象のためのメタデータ

検索対象ベクトルを生成するために，検索対象を特徴づけたメタデータである．検索対象の特徴を表す語で特徴づける．特徴づけに用いられた各特徴の関連は，検索空間生成のためのメタデータで記述される．

(3) 問い合わせのためのメタデータ

問い合わせベクトルを生成するために，問い合わせを特徴づけたメタデータである．特徴づけに用いられた各特徴の関連は検索空間のためのメタデータで記述される．

特徴づけに用いられる語

特徴づけに用いられる語	語1	語2	語3	…
語1	1	0	-1	…
語2	0	1	0	…
語3	-1	0	1	…
⋮	⋮	⋮	⋮	⋱

図 2 語同士の関係を直接記述するメタデータの例．
Fig. 2 An example of the metadata representing relation between each word directly.

本稿では，上記 3 つのメタデータのうち「検索空間生成のためのメタデータ」に注目する．このメタデータは特徴として用いる語と語の関連を表すものである．特定分野の検索方式の実現において，このメタデータを正確にかつ容易抽出することが非常に重要となる．さらに，様々な情報源から取り出された「検索空間生成のためのメタデータ」を統合し，新しい検索空間が形成できるならば，より多角的な検索が可能になると考えられる．本稿ではこのメタデータ群の統合方式について言及する．

3. 検索空間生成のためのメタデータ抽出方式

本節では，検索空間生成のためのメタデータ抽出方式について述べる．

3.1 検索空間生成のためのメタデータの形式

2 章のデータ行列の形式から検索空間生成のためのメタデータとして大きくわけて 2 種類の形式が考えられる．

(1) 語同士の関係を直接記述するメタデータ

各語同士の関係を一つずつ重み付けしていく方法である．具体的な例としては図 2 に示す．

本メタデータの形式は最初から実対称行列を構成できる．主に専門家によって手作業で特徴づける場合において付与される場合この形式に

特徴づけに用いられる語

語を関連付ける事柄（基本データ）

	語1	語2	語3	…
データ1	0	1	0	…
データ2	1	-1	0	…
データ3	-1	0	1	…
⋮	⋮	⋮	⋮	⋱

図 3 語同士の関係を間接的に記述するメタデータの例.
Fig. 3 An example of the metadata representing relation between each word indirectly.

なることが多い.

(2) 語同士の関係を間接的に記述するメタデータ

各語同士の関係を直接記述するのではなく，ある基本データを介して重み付けしていく方法である．具体的な例としては図 3 に示す.

本メタデータの形式は，ある既存の特定分野の内容が書かれた資料から構成する場合に多い．この方法における行ベクトルを基本データと呼ぶ．基本データを何に設定するかによって，どのような関係を表すメタデータなのかが違ってくる．

3.2 検索空間生成のためのメタデータ抽出方法

本節では，検索空間生成のためのメタデータの抽出方法についてこれまでの方式の長所，短所を考察しながら示す．具体的には，検索対象のメタデータを検索空間生成のためのメタデータとして用いる方法，辞書を用いる方法，そして我々の研究を進めている書籍の索引部を用いた方式を示す.

3.2.1 検索対象のメタデータを検索空間生成のためのメタデータとして用いる方法

この方法は [2),3)]，主に LSI(Latent Semantic Indexing)[2),3)] で採用されており，同検索対象データにメタデータとして付与されている語同士は関係があるとみなし，検索対象のメタデータをそのまま検索空間生成に利用する．この方法は，検索空間生成のために，新たなメタデータを抽出，用意することなく容易に少ないコストで検索空間の生成ができる.

しかしながら，検索対象が増えるごとに空間の更新という重い処理を行わないといけない上，検索対象のメタデータとして全く関連しない語同士が付与されていた場合，そもそも検索対象のメタデータがうまく付与されていないものが多い場合，検索空間全体の精度にも影響するという問題点がある．分野によってメタデータの付与の仕方に差がある場合，関連のある語同士が関連しない場合もある．この方法は分野が限られていない一般的な検索対象に対して有効である.

3.2.2 英英辞典や用語辞典を用いる方法

Longman Dictionary of Contemporary English(以下，Longman)[6)] という英英辞典を用いる方法がある [5)]．また文献 [8),9)] では，それぞれの特定分野の用語辞典を用いて，特定分野を対象とした意味を計量可能な検索空間を生成する方式が提案されている．この方法は主に意味の数学モデル [4),5),7)] で採用されている．これらの方式 [5),8),9)] は，説明される言葉 (見出し語) を語義文中に用いられている言葉 (特徴語) によって特徴づけすることによって，データ行列を作成し，検索空間を生成している．これにより，それぞれの分野における質の高い検索空間を実現できる.

これらの特定分野における用語辞典によるメタデータ空間生成方式 [8),9)] では，見出し語の特徴をよく表す特徴語を抽出するために，語義文中の語以外で言い換えて特徴づけしたり，特徴語を選別するため，必ず専門家による作業が必要となる．このため，自動化が難しく高い専門性を必要とする.

また，このような用語辞典を用いた方式 [8),9)] の適用が難しい用語辞典が存在する．その例として,「岩波数学辞典 第 3 版」[12)] が該当する．本方式が適用できなかったり，例えば趣味を扱う語においては，用語辞典の存在しない場合が多いと考えられる．ただし，対象の分野に適した辞典が存在する場合は非常に有効な方法である.

3.2.3 書籍の索引部を用いた方式

本方式 [10)] は，対象とする特定分野について書かれた教科書や参考書レベルの書籍を対象とし，索引を用いて検索空間を生成する．具体的には，次の流れで実現する.

(1) 初期データ行列の設定

まず，対象とする特定分野について書かれた書籍の索引を参照する．索引に出現する語を特徴語とみなし，索引情報から各ページ番号を用いて特徴づける.

$$\mathbf{p}_i = (f_{i1}, f_{i2}, \cdots, f_{in}) \tag{5}$$

ここで i はページ番号，f_{ik} は特徴語に対応したページ番号について特徴づけた値である．特徴づける f_{ik} の値は，以下のように決定される.

- 索引中で特徴語がそのページ番号を参照している場合：”1”
- 索引中で特徴語がそのページ番号を参照してない場合：”0”

以上から，$\mathbf{p}_i$ を用いて，$(\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_m)^T$ とすることによって，m 行 n 列の初期データ行列 M_0 を作成する.

(2) 初期データ行列 M_0 の修正によるデータ行列 M の生成

(1) で作成した初期データ行列 M_0 には，ページ番号と語の関係を表す行列となっており，ページ同士の関係が反映されていない．そのため，ある概念が複数ページにわたって書かれている場合，索引に記述されているキーワードとして表される語とページ番号の関係だけでは表現しきれず，精度を悪化させる原因となりうる．初期データ行列 M_0 にページ同士の関係を反映するように修正してデータ行列 M を生成する．まず，章，節の番号を特徴語として初期データ行列 M_0 を修正，追加する．章，節番号について該当ページを全て”1”，それ以外のページを”0”と特徴づける．例えば，23 ページが 2 章 3 節に該当する場合,「2」,「2-3」を特徴語として，23 ページの「2」,「2-3」に”1”と特徴づける．以上により，m 行 $n+\alpha$ 列のデータ行列 M を生成できる．ここで，α は章，節番号を特徴として付け加えた分である．

書籍によって検索空間を生成することから，容易に，専門知識を必要せず，少ないコストで，検索空間の生成ができる．また書籍は新しい確実な専門知識を辞典よりも早く入手することができる．しかしながら，一般的に書籍 1 冊の索引に収録されている語数は数百から数千で少なく，検索空間上で用いることができる語彙数が少なくなってしまうという問題点がある．

大きく 3 つの方法を示したが，これらを組み合わせて他の空間と統合することが出来れば，それぞれの問題点が解決し，より語同士の関係に合致したベクトル空間モデルによる検索システムが実現できると考えられる．次章で，その検索空間の統合方式について述べる．

4. 検索空間統合の実現方式とその検討

本節では，3 章で生成されたデータ行列を複数組み合わせることにより，より広い分野を網羅する検索空間を生成する方式について示す．

4.1 検索空間統合の実現方式

ここでは，索引から抽出されたデータ行列 M_A と他の情報から抽出されたデータ行列 M_B を対象として統合したデータ行列 M_C の生成方式を示す．本方式の全体図を図 4 に示す．なお，これは図 4 に示す通り 2 者のみの統合方式ではなく，複数統合することが可能な方式である．

(1) M_A と M_B の特徴群の統合

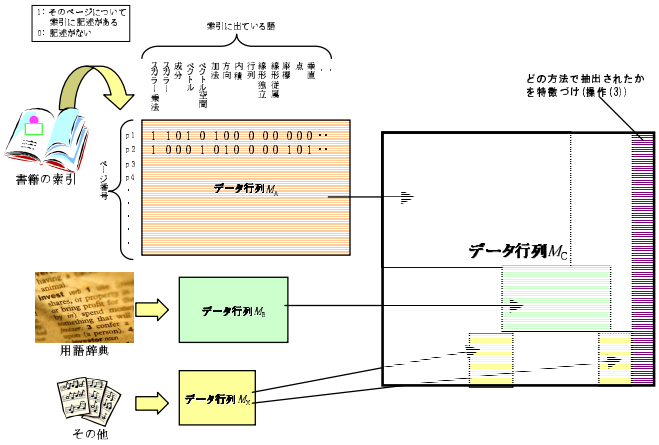

図 4 検索空間統合方式の概要．
Fig. 4 An outline of an integration method of retrival spaces.

M_A，M_B 間において，それぞれの列要素である特徴群を合成し，特徴語の重複を除く．この集合を，統合したデータ行列 M_C の特徴群とする．

(2) M_A と M_B の基本データ群の統合

行要素である基本データ群の重複を除く，基本データ群の統合がある．しかしながら，本方式は異種の情報源語と語の関係を示しており，この集合を，統合したデータ行列 M_C の基本データ群とする．なお，このような重複は見られない場合もある．

(3) 統合されたデータ行列の修正

各データ行列がどのように抽出されたか (例えば，索引によって抽出，辞書から抽出など) ということも語と語の関連を計量することにおいて重要な要因になりうる．よって，(1) で作成したデータ行列がどのように抽出されたかの関係を反映するように修正してデータ行列 M_C を完成させる．まず，抽出元 ID(一意であればなんでもよい) を特徴として (1) で作成したデータ行列を修正，追加する．その特徴において該当する基本データを全て”1”，それ以外のページを”0”と特徴付ける．

ここで，(1) 及び (2) においての，M_A に付与されている重みの値と M_B に付与されている重みの値の統合方法を示す．この方法は，各データ行列の性質などによって異なる可能性があるが，本稿では以下のように仮に設定する．

- それぞれの付与された値が両方とも同じ符号の場合

それぞれの値を掛け合わせた値を統合した新しい値とする．

- それぞれの付与された値が両方とも 0 の場合
  統合した値も 0 とする．
- それぞれの付与された値が異なる符合の場合
  本方式では，正である方の値をとることとする．これは，実際本当に正か負かという判断をするとなると，専門家による判断によるのが一番である．しかしながら，自動化を考えるのであれば人手による判断はなるべく避けたい．ここで，肯定の意味であれ否定の意味であれ，その要素に非ゼロの値が付与されているということは，何らかの関係を持っているということがいえるため，人手を使わず自動化して行う場合は，正の値を統合後の値として用いる．

以上により，統合したデータ行列 M_C が生成できる．上記の例は 2 つのデータ行列を拡張統合実現する方式であるが，3 つ以上についても，同様の方式で拡張統合可能である．その際データ行列の統合順序に拡張統合結果は依存しない．

4.2 検索空間統合方式の検討

本節では,「日経パソコン用語辞典 2004 CD-ROM 版」(以下，パソコン用語辞典)[13] と初級システムアドミニストレータの教科書である「情報処理教科書システムアドミニストレータ平成 15 年度版【春期】」(以下，教科書)[14] の一部を用いて小さなデータで異種の情報群から生成されたもの同士の検索空間統合の検討する．

まず，パソコン用語辞典の「DBMS」「データベース管理システム」「アプリケーション」という見出し語の語義文の語をメタデータとして抽出した．30 個の語で構成されている．これのみで検索空間を生成すると，3 次元の正規直交空間となる．また，教科書の 92 ページから 95 ページ，つまり「DBMS」の記述のあるページに関する索引のみを用いてメタデータを抽出した．17 個の語で構成されている．なお，ここでは目次，つまり章節は考慮していない．これのみで検索空間を生成すると，4 次元の正規直交空間となる．さらに，これら 2 つを統合した検索空間を生成した．これは 44 個の語で構成されており，7 次元の正規直交空間となる．これらを意味の数学モデルの計量系を用いて「データベース」,「DBMS」を問い合わせとしてそれぞれの空間においてどの語が近いと検索されるかを考察した．

問い合わせとして「データベース」を与えたときのそれぞれの上位 5 位の結果を表 1,2,3 に示す．パソコン用語辞典のみの検索空間の場合,「データベース」自身の他,「管理」,「ソフト」,「アプリケーション」,「挿入」が上位に検索されているが,「DBMS」などの最も関連すると思われる語が上位に検索されていない．また，教科書のみの検索空間の場合においても「データベース」自身以外関連を見出せない．それに対して，この 2 つを統合した検索空間の場合,「データベース」自身の他,「リレーショナルデータベース管理システム」「RDBMS」「DBMS」「データベース管理システム」が上位に検索されている．これより，各異種情報源による検索空間の結果が思わしくない場合でも，それらの空間を統合し新しい検索空間とすることで，語同士の関係の計量結果が良くなることが確認できる．

表 1 パソコン用語辞典のみの場合 (コンテクスト：データベース)
Table 1 Case of utilizing a dictionary (Context:データベース).

管理	0.910181
ソフト	0.910181
データベース	0.910181
アプリケーション	0.910181
挿入	0.694363

表 2 教科書の索引のみの場合 (コンテクスト：データベース)
Table 2 Case of utilizing a index of book (Context:データベース).

データベース	1.000000

以下検索されない．

表 3 統合した場合 (コンテクスト：データベース)
Table 3 Case of an integration of 2 data (Context:データベース).

データベース	0.994548
リレーショナルデータベース管理システム	0.054883
RDBMS	0.054883
DBMS	0.054190
データベース管理システム	0.054190

問い合わせとして「DBMS」を与えたときのそれぞれの上位 5 位の結果を表 4,5,6 に示す．パソコン用語辞典のみの検索空間の場合，上位に「DBMS」と「データベース管理システム」が検索されいる．教科書のみの検索空間の場合,「DBMS」「データベース管理システム」「RDBMS」「リレーショナルデータベース管理システム」以外関連を見出せてないが，関連のある 4 語が検索されている．この 2 つを統合した検索空間の場合も「DBMS」「データベース管理システム」「RDBMS」「リレーショナルデータベース管理システム」が上位に検索され，かつ他の語も関連を見出している．これより，各異種情報源による検索空間の結果が良い場合，それらの空間を統合し新しい検索空間と

表 4 パソコン用語辞典のみの場合 (コンテクスト：DBMS)
Table 4 Case of utilizing a dictionary (Context:DBMS).

DBMS	0.971908
データベース管理システム	0.971908
変換	0.235240
ツール	0.235240
データベースソフト	0.235240

表 5 教科書の索引のみの場合 (コンテクスト：DBMS)
Table 5 Case of utilizing a index of book (Context:DBMS).

DBMS	1.000000
RDBMS	1.000000
データベース管理システム	1.000000
リレーショナルデータベース管理システム	1.000000

以下検索されない.

表 6 統合した場合 (コンテクスト：DBMS)
Table 6 Case of an integration of 2 data (Context:DBMS).

DBMS	0.892732
データベース管理システム	0.892732
リレーショナルデータベース管理システム	0.579711
RDBMS	0.579711
ツール	0.256794

することによって精度が下がらないと予想される.

これらの検討により，書籍の索引によって生成された検索空間と辞書によって生成された検索空間の統合の有効性は十分にあると予想される．なお，実際の実用レベルでの検索空間による検証実験は今後の課題である.

5. おわりに

本稿では，書籍の索引部を用いた検索空間と他の情報元による検索空間との統合方式のアイデアを示した．本方式が実現されることにより，メタデータ空間を生成したい対象となる特定分野のことについて書かれた複数書籍を準備し，索引を参照することで，その特定分野を網羅する検索空間に既存の空間を統合することによる新しい検索空間生成することが可能となると考えられる.

本方式により，これまで実現できなかった特定分野にも，ベクトル空間モデルを用いた検索機構の導入が容易に可能になると考えられる．さらに，特定分野の専門家がその関心に基づく質の高い情報群を対象とした利用者が意図する情報を獲得する効率が高い検索が各分野でそれぞれ実現されることにより，各特定分野のネットワーク・コミュニティの活動のパフォーマンスを増大させることが可能であると考えられる.

今後の課題として，本方式の具体的な定量的な評価，本方式の他分野を対象とした検索方式への適用，大規模検索空間の生成における固有値分解，特異値分解の効率化が挙げられる.

参 考 文 献

1) R.Baeza-Yates, B.Ribeiro-Neto, “Modern Information Retrieval,” Addison Wesley, (1999).
2) M.W.Berry, S.T.Dumais, and G.W.O’Brien, “Using linear algebra for intelligent information retrieval,” SIAM Review Vol. 37, No.4, pp.573-595 (1995).
3) S.Deerwester, S.T.Dumais, G.W.Furnas, T.K.Landauer, and R.Harshman, “Indexing by Latent Semantic Analysis, ” Journal of the American Society for Information Science, Vol. 41, No. 6, pp.391-407 (1990).
4) T.Kitagawa, and Y.Kiyoki, “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(1993).
5) Y.Kiyoki, T. Kitagawa, and T.Hayama, “A Metadatabase System for Semantic Image Search by a Mathematical Model of Meaning,” Multimedia Data Management – using metadata to integrate and apply digital media –, McGrawHill, A. Sheth and W. Klas(editors), Chapter 7 (1998).
6) Longman Dictionary of Contemporary English, Longman (1987).
7) 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).
8) 宮川祥子，清木康, “特定分野ドキュメントを対象とした意味的連想検索のためのメタデータ空間生成方式 ,” 情報処理学会論文誌：データベース, Vol.40, No.SIG5(TOD2), pp.15-27,(1999).
9) 河本穣，清木康，吉田尚史，藤島清太郎，相磯貞和, “医療分野ドキュメント群を対象とした意味的連想検索空間の実現方式 ,” 日本データベース学会 Letters, Vol.1, No.2, pp.12-15,(2003).
10) 中西 崇文, 岸本 貞弥, 櫻井 鉄也, 北川 高嗣: “特定分野を対象とした連想検索のための書籍の索引部を用いたメタデータ空間生成方式,” 電子情報通信学会論文誌,VOL.J88-D1, No.4, pp.840-851, (2005).
11) 石原冴子，清木康, “異分野データベース群を対象とした意味的検索空間統合方式とその実現 ,” 情報処理学会論文誌：データベース, Vol.43, No.SIG5(TOD15), pp.15-27,(2002).

12) 岩波 数学辞典 第 3 版，岩波書店，(1985).
13) “日経パソコン用語辞典 2004 CD-ROM 版, ” 日経 BP 社, (2003).
14) 工房 mana: “情報処理教科書システムアドミニストレータ平成 15 年度版【春期】,” 翔泳社, (2002).