DEWS2004 I-9-02

複数の書籍の索引部を用いたメタデータ空間拡張統合方式

中西　崇文† 　岸本　貞弥†† 　櫻井　鉄也††† 　北川　高嗣†††

† 筑波大学大学院 システム情報工学研究科　〒 305–8573 茨城県つくば市天王台 1–1–1 数値解析研究室
†† 筑波大学大学院 理工学研究科　〒 305–8573 茨城県つくば市天王台 1–1–1 数値解析研究室
††† 筑波大学 電子・情報工学系　〒 305–8573 茨城県つくば市天王台 1–1–1
E-mail: †{takafumi,kishimoto}@nalab.is.tsukuba.ac.jp, ††{sakurai,takashi}@is.tsukuba.ac.jp

あらまし　本稿では，複数の書籍の索引部を用いたメタデータ空間拡張統合方式を示す．本方式は，それぞれの書籍の索引部を用いて生成された各データ行列を対象に，どちらの書籍の索引にも用いられている共通の語を用いて，単語間の関連をともなった検索空間であるメタデータ空間の拡張，統合を実現する．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．本方式は，複数の書籍を用いることにより，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．本稿では，本方式を意味の数学モデルにおける単語間関連連想検索に適用する際の実現方式を示し，本方式の有効性を確認する．
キーワード　メタデータ空間生成，検索空間，空間統合，意味の数学モデル，単語間関連連想検索

An integration and extension method of metadata spaces from Indexes of Documents

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

† Graduate School of Systems and Information Engineering, University of Tsukuba
†† Master's Program in Science and Engineering, University of Tsukuba
††† Institute of Information Sciences and Electronics, University of Tsukuba
E-mail: †{takafumi,kishimoto}@nalab.is.tsukuba.ac.jp, ††{sakurai,takashi}@is.tsukuba.ac.jp

Abstract　In this paper, we present an integration and extension method of metadata spaces utilizing the index part of two or more books. This method make it possible to integrate metadata spaces based on the relation between words by using common terms in indexes of documents. The metadata space by the index of a book is in the tendency whose number of words which can be expressed decreases with the character of books. The problem is solved when two or more books are used for this method. It is thought that the metadata space established method by the index part of the books containing this method is applicable to mediadata and document search for brooad fields, such as a field of not only a scientific field but a hobby In theis paper, we also present an implementation method for applying our method to words related associative search. We clatify effectiveness of our method by several experiments.
Key words　Establishment of a metadata space, Retrieval space, Space integration, Mathmatical model of meaning, Words related associative search

1. ま え が き

現在，様々な特定分野の専門家がその関心に基づく質の高い情報を他の利用者に提供するシステムへのニーズが高まっている．既に，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散在しつつある．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，高度な検索方式と知識の発掘方式が重要となっている．

文献 [1] [2] [3] で，言葉と言葉の関係の計量による検索機構として，意味の数学モデルを提案している．これは, 単語群を文脈として解釈する機構により，言葉と言葉，あるいは，言葉と検索対象のメディアデータ，ドキュメント間を文脈に応じて動的に計算することを可能とする．意味の数学モデルでは，検索対象をベクトル化し，メタデータ空間と呼ばれる空間に写像する．さらに，それらのベクトルをメタデータ空間の部分空間

に射影して計量することにより，文脈に応じた連想検索を実現している．

意味の数学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，その特定分野を表現するためのメタデータ空間を作成する必要がある．意味の数学モデルでは，メタデータ空間を基本データとよばれる特徴付きベクトルの集合であるデータ行列から生成する．各特定分野の特徴を反映したメタデータ空間を生成するためには，このデータ行列を適切な方法で作成する必要があり，その生成方式が問題となる．

特定分野のデータ行列の生成方式として，これまで文献[4][5]で，辞書や用語辞典を用いて生成する方式が提案されている．これらの方式によって，特定分野を対象とした意味を計量するための良質なメタデータ空間生成を可能とし，意味的連想検索を実現している．しかしながら，これらの方式は，一定の条件を満たす，辞書や用語辞典があることを前提としており，これらの辞書や用語辞典がない特定分野について，実現が困難であることが問題であった．また，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[4][5]では，必ず人間による作業が必要となる．例えば，文献[4]の方式では，見出し語の特徴づけとして，その見出し語の語義文で用いられている特徴語以外で，特徴づけを行わなければいけないケースを示している．また，文献[5]の方式では，特徴づけのための特徴語を用語辞典中の説明のために用いられる頻度と，多くの概念を特徴付けるために，必要と思われる普遍性の高い語であるかによって，選別をして，見出し語に特徴づけしている．これらの作業はその分野の専門家レベルの人間でないと難しい作業である．

このことから，これらの方式[4][5]では，自動化は不可能であり，多大な時間を要する．

これまで我々は，特定分野の書籍の索引部を用いて単語の関連を計量する専門分野を対象としたメタデータ空間を生成する方式[6]について研究を進めてきた．この方式は，書籍によってメタデータ空間を生成することから，適切な辞書や用語辞典がない分野について，有効である．また，この方式は，容易に，専門知識を必要せず，完全自動化可能なメタデータ空間生成方式である．そのことから，用語辞典による方式[2][4][5]の代替方式としても有効である．これらから，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．

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あり，用語辞典に収録されている語数より少ないため，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．一方，教科書などの書籍は，辞典が扱う分野に比べて，より細かく狭い傾向がある．例えば，辞典ではIT分野など広い範囲で書かれているのに対し，教科書などの書籍は，データベース，ネットワークなど，より細分化された範囲で書かれている傾向にある．細分化された専門的な分野における語での計量で十分な場合は，1冊の書籍を用いて，少ない手間で空間を生成し，利用できる．より語彙数を増やしたい場合に，索引によるメタデータ空間生成方式における拡張統合方式が重要である．

空間の統合に関して，意味の数学モデルを用いた意味的連想検索を対象とした，メタデータ空間の統合方式として，文献[7]が実現されている．この方式は，辞書や用語辞典で生成されたデータ行列を対象として，異分野の統合を目的として統合する方式を示されている．

書籍の索引部を用いて生成されたデータ行列を対象として，空間拡張統合する方式は実現されれば，書籍の索引部で作成するデータ行列の問題である，生成できたメタデータ空間が扱うことのできる語の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を回避できると考えられる．

本稿では，複数の書籍の索引部を用いたメタデータ空間拡張統合方式について示す．本方式は，複数の書籍の索引部を用いて，その索引部を組み合わせて，比較的扱う分野の範囲が大きい空間を作成する方式である．本方式により，文献[6]の方式でのその空間が扱うことのできる語の範囲が狭くなる問題を回避する．本方式によって，より広範囲の語と語の関連性による，意味の数学モデルによる単語間関連連想検索が実現される．

ここで，意味の数学モデルを用いた連想検索方式は，文献[8989]に代表される，LSIと呼ばれる多変量解析による空間生成を用いた検索手法とは次の点で本質的に異なる．意味の数学モデルを用いた連想検索方式では，直交空間における部分空間選択を行う演算を定義し，その演算により，言葉の意味的関係を，文脈，すなわち与えられた検索要求に基づいて選択された部分空間に応じて，解釈するという機構を実現している．意味の数学モデルとLSIの違いについて，詳細は，文献[10]で報告されている．

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書籍を複数準備し，索引を参照することで，その特定分野を網羅するメタデータ空間を，従来方式に比べ，容易に，かつ，専門知識がなくても生成が可能となる．このことから，これまで，実現できなかった特定分野にも，語と語の関連による連想検索の導入が容易に可能になると考えられる．これにより，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獲得する効率が高い検索が各分野でそれぞれ実現され，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の第一歩となりうる．

本稿では，IT分野を対象とし，複数の書籍の索引部を用いてメタデータ空間生成し，意味の数学モデルを適用した単語間関連連想検索を実現する．また，IT分野における画像メディアデータを対象として単語間関連画像メディアデータ検索を実現し，実験により，本方式の有効性を検証する．

2. 書籍の索引部を用いたデータ行列生成概要

本節では，書籍の索引部を用いたメタデータ空間を生成するためのデータ行列の生成方式の概要を示す．本方式は，近辺の単語同士は相関が高いとして，ページ番号を場所情報として特徴付けたデータ行列を生成する方式である．

詳細は，文献[6]で示されている，

（1） 初期データ行列の設定

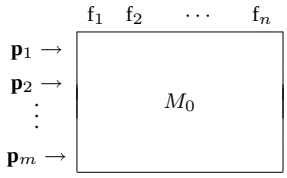

図 1　初期データ行列 M_0 によるメタデータの表現.

Fig. 1　Metadata represented in first data matrix M.

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付ける．

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

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付けた値である．特徴付ける 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，図 1 のような m 行 n 列の初期データ行列 M_0 を作成する．

（2） **初期データ行列の修正によるデータ行列の生成**

(1) で作成した初期データ行列 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_1 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”と特徴付ける．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”と特徴付ける．

以上により，m 行 $n+\alpha$ 列のデータ行列 M_1 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．

3. 書籍の索引部を対象としたメタデータ空間拡張統合方式

本節では，対象となる分野の書籍を複数用意し，それぞれを用いて 2. 節で生成された，複数のデータ行列を組み合わせることにより，対象分野を網羅するような空間を生成する方式について示す．ここでは，書籍 1 の索引から生成されたデータ行列 M_1，書籍 2 の索引から生成されたデータ行列 M_2 を対象として統合したデータ行列 M の生成方式を示す.

（1） M_A **と** M_B **の特徴群の統合**

M_A，M_B 間において，それぞれの特徴群を合成し，特徴語の重複を除く．この集合を，統合したデータ行列 M の特徴群とする．この概要図を図 2 に示す.

なお，文献 [7] では，辞書や用語辞典の見出し語の重複を除く，基本データ群の統合がある．索引部の場合，基本データ群はページ番号に相当する．ここで，同じページ番号であったとしても，別の書籍であれば，場所情報としてまったく異なるものであると言える．そのため，文献 [7] のベクトル要素の統合も，図 2 に示すとおりない.

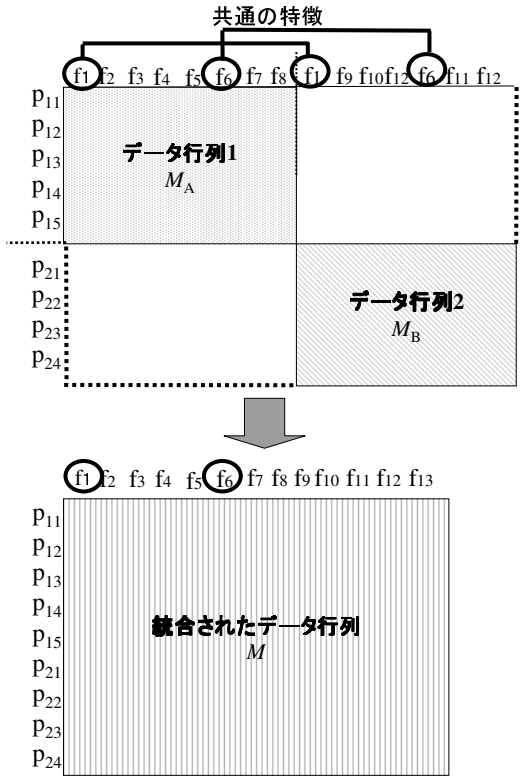

図 2　特徴の統合.

Fig. 2　Integration for features.

（2） **統合されたデータ行列の修正**

書籍は一般的にある分野の特定の部分を説明するものである．同様の分野の書籍を見比べても，著者や対象とする読者によって書かれ方や使用される語が全く異なる場合がある．また，例えば「IT 分野」と特定した場合，辞書は IT 分野の辞書として，1 冊にまとめられている場合が多いが，書籍の場合は「データ工学」,「自然言語処理」,「プログラミング」など多岐に亘る．このことから，どの書籍に記述されていたかということも言葉と言葉の関連を計量することにおいて重要な要因になりうる．これらの書籍の索引を組み合わせる場合，書籍同士の関係を反映する必要がある.

よって，(1) で作成したデータ行列に書籍同士の関係を反映するように修正してデータ行列 M を完成させる.

まず，本の ID(一意であればなんでもよい) を特徴語として (1) で作成したデータ行列を修正，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”と特徴付ける．例えば，2 つの書籍を対象とする場合,「書籍 1」「書籍 2」を特徴語として，書籍 1 の索引のページ番号の場合「書籍 1」に”1”，書籍 2 の索引のページ番号の場合「書籍 2」に”1”とそれぞれ特徴付ける.

以上により，統合したデータ行列 M 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. 意味の数学モデルへの適用

本節では，3. 節で拡張統合されたメタデータ空間を意味の数学モデルに適用することにより，単語間関連連想検索の実現方法を示す．意味の数学モデルの詳細は，文献 [1] [2] [3] に示している．

（1） **メタデータ空間 $\mathcal{MDS}$ の生成**

データ行列 M からメタデータ空間 $\mathcal{MDS}$ を生成する．メタデータ空間は，データ行列 M の相関行列 M^TM を固有値分解し，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．

（2） **検索対象データのメタデータをメタデータ空間へ写像**

メタデータ空間へ検索対象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，検索対象データが同じ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関係を空間上での語と語の関係として計算することが可能となる．

検索対象データ D には，メタデータとして t 個の語 $o_1, o_2, \cdots, o_t$ が以下のように付与されていることを前提としている．

$$D = \{\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_t\}. \tag{2}$$

ここで，各印象語 $\mathbf{o}_i$ は，データ行列の特徴語と同一の特徴を用いて表現される特徴付ベクトルである．

$$\mathbf{o}_i = (o_{i1}, o_{i2}, \cdots, o_{in}) \tag{3}$$

各検索対象データは，メタデータとして付与されている t 個の語が以下のように合成され，検索対象データベクトル $\mathbf{d}$ を形成する．

$$\begin{aligned}\mathbf{d} &= \bigoplus_{i=1}^{t} \mathbf{o}_i \\ &:= (\mathrm{sign}(o_{\ell_1 1}) \max_{1 \leqq i \leqq t} |o_{i1}|, \\ &\qquad \mathrm{sign}(o_{\ell_2 2}) \max_{1 \leqq i \leqq t} |o_{i2}|, \\ &\qquad \cdots, \mathrm{sign}(o_{\ell_n n}) \max_{1 \leqq i \leqq t} |o_{in}|).\end{aligned} \tag{4}$$

この和演算子 $\bigoplus_{i=1}^{t}$ は，t 個のベクトルから各基底に対して絶対値最大の成分を選ぶ演算子である．ここで sign(a) は，“a” の符号（正，負）を表す．また，$l_k (k = 1, \ldots, t)$ は，特徴が最大となる印象語を示す指標であり，次のように定義する．

$$\max_{1 \leqq i \leqq t} |o_{ik}| = |o_{\ell_k k}|. \tag{5}$$

これにより検索対象データのメタデータがデータ行列の特徴語と同一の特徴を用いて表現される．検索対象データベクトル $\mathbf{d}$ をメタデータ空間へ写像する．この写像は，検索対象データベクトル $\mathbf{d}$ をメタデータ空間内でフーリエ展開し，フーリエ係数を求める．

（3） **メタデータ空間の部分空間の選択と相関の定量化**

検索者が与える単語の集合をコンテクストと呼ぶ．コンテクストを用いてメタデータ空間に各単語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を持つ軸からなる部分空間が選択される．選択されたメタデータ空間の部分空間において，検索対象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．これにより検索者が与えた検索語と各ドキュメントデータとの相関の強さを定量化する．この部分空間における検索結果は，各検索対象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．

5. 実　　験

本方式の有効性を検証するため，IT 分野を対象として拡張統合方式をもちいて生成されたメタデータ空間について，検証実験を行った．「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】」[11] の索引部で生成した空間 (以下，IT 分野の空間) と，「データベースシステム」[12] の索引部で生成した空間 (以下，データベース関連の空間) を統合することにより，空間を拡張する．

予備実験では，「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】」[11] の索引部で生成した空間と，「データベースシステム」[12] の索引部で生成した空間をそれぞれについて個々の検索精度を検証する．

実験 A では，2 つを拡張統合した空間での単語間関連連想検索による検索精度を検証する．

5.1 実 験 環 境

「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】」[11] の索引に出現するすべて語を対象として，3. 節 (2) に記述がある，索引部と目次の情報を入れたデータ行列から作成した 170 次元のメタデータ空間を生成した．

また，「データベースシステム」[12] の索引に出現するすべて語を対象として，3. 節 (2) に記述がある，索引部と目次の情報を入れたデータ行列から作成した 169 次元のメタデータ空間を生成した．

さらに，この二つを統合することにより，339 次元のメタデータ空間に拡張した．

5.2 予 備 実 験

まずは，「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】」[11] の索引部で生成した空間である IT 分野の空間と，「データベースシステム」[12] の索引部で生成した空間であるデータベース関連の空間をそれぞれについて個々の検索精度を検証する．

5.2.1 実 験 方 法

本来は，4. 節で示すように，検索対象となる検索対象データに対して，式 (2) のように 1 語以上の単語をメタデータとして付与し，単語間関連連想検索を行う．本実験では，生成したメタデータ空間の性質を考察するための実験のため，それぞれのメタデータ空間が扱うことができる用語を検索対象として実験を行った．

なお，語を検索することは，全ての検索対象データに対して，それぞれ異なる 1 語のみからなるメタデータを付与し，その検索対象データを検索していることと等価である．また，メタ

表 1　実験結果 1-1(コンテクスト：TCP/IP).
Table 1　Experimental results 1-1 (Context:TCP/IP).

語	相関量
TCP/IP	0.828350
OSI 参照モデル	0.322581
IP アドレス	0.223156
インターネットサービスプロバイダ	0.175884
DHCP サーバ	0.175884
プロバイダ	0.175884
FTP	0.167232
Telnet	0.167232
Gopher	0.167232
SMTP	0.148756

表 2　実験結果 1-2(コンテクスト：5 大装置).
Table 2　Experimental results 1-2 (Context:5 大装置).

語	相関量
マザーボード	0.759374
制御装置	0.759374
5 大装置	0.759374
出力装置	0.676300
記憶装置	0.500466
演算装置	0.416007
入力装置	0.414988
パソコン	0.313336
PC カード	0.278557
PCMCIA	0.278557

表 3　実験結果 1-3(コンテクスト：データベース).
Table 3　Experimental results 1-3 (Context:データベース).

語	相関量
データベースシステム	0.886946
DBS	0.886946
データベース	0.886946
DB	0.886946
DBMS	0.174252
データベース管理システム	0.174252
データサブ言語	0.160500
問い合わせ言語	0.160500
質問言語	0.160500
データベース管理者	0.160500

表 4　実験結果 1-4(コンテクスト：集合関数).
Table 4　Experimental results 1-4 (Context:集合関数).

語	相関量
AVG	0.964398
SUM	0.964398
COUNT	0.964398
グループ表	0.964398
集合関数	0.964398
GROUP_BY 句	0.964398
MIN	0.964398
MAX	0.964398
PCMCIA	0.079548
SQL2	0.079548

データは，1 語であったとしても，複数個の語であったとしても，式 (3),(4) より，データ行列と同じ特徴によって表され，メタデータ空間に写像される．そのため，メタデータ空間の性能評価としては，語を検索対象とすることで十分であると言える．

それぞれの空間において，問い合わせを発行した．実験結果において，上位 10 件について，考察をおこなう．

5.2.2　実 験 結 果

まず，IT 分野の空間による検索結果を，表 1, 表 2 に示す．

表 1 では，コンテクストとして「TCP/IP」を与えている．その結果，関連のある「OSI 参照モデル」「IP アドレス」や,「FTP」「Telnet」「SMTP」などのプロトコルが上位に出力されている．

表 2 では，コンテクストとして「5 大装置」を与えている．その結果,「5 大装置」である「制御装置」「出力装置」「記憶装置」「演算装置」「入力装置」がそれぞれ，2 位，4 位，5 位，6 位，7 位に出力されている．

これらによって，ほぼ関連した語が出力されていることから，IT 分野のメタデータ空間において，単語間関連連想検索が適切に行われていることがわかる．

さらに，データベース関連の空間による検索結果を，表 3, 表 4 に示す．

表 3 では，コンテクストとして「データベース」を与えている．その結果，関連のある「データベースシステム」「DBS」「DB」が同位 1 位,「DBMS」「データベース管理システム」が同位 5 位に出力されている．

表 4 では，コンテクストとして「集合関数」を与えている．その結果,「集合関数」である「AVG」「SUM」「COUNT」「MAX」「MIN」が同位 1 位に出力されている．

これらによって，ほぼ関連した語が出力されていることから，データベース分野のメタデータ空間において，単語間関連連想検索が適切に行われていることがわかる．

しかしながら，相関量の値が同じ値の語が多数出力されることから，ひとつの書籍の索引部で生成した空間の場合，関連を表す様相が少ないことが考えられる．このことからも，複数の書籍を用いて拡張統合行う重要性を示している．

5.3　実　験　A

5.3.1　実 験 方 法

「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】」[11] の索引部で生成した空間と,「データベースシステム」[12] の索引部で生成したデータ行列を組み合わせることで，2 つを拡張統合した空間での単語間関連連想検索による検索精度を検証を行った．

IT 分野の空間，データベース関連の空間，拡張空間の 3 空間に対して，同一のコンテクストを発行し，それぞれの検索結果を比較する．予備実験と同様，メタデータ空間が扱うことができる用語を検索対象として実験を行った．

5.3.2　実 験 結 果

コンテクスト「データベース管理システム」で検索した場合,「ジャーナル」で検索した場合,「ファイル」で検索した場合を表 5，表 6，表 7 にそれぞれ示す．

コンテクスト「データベース管理システム」の場合，表 5 から，IT 分野の空間では,「リレーショナルデータベース管理システム」「RDBMS」「DBMS」が「データベース管理システム」と同位で 1 位に出力されている．データベース関連の空間では,

表 5　実験結果 A-1(コンテクスト：データベース管理システム).

Table 5　Experimental results A-1 (Context:データベース管理システム.

拡張空間		IT 分野の空間		DB 分野の空間	
データベース管理システム	0.725401	リレーショナルデータベース管理システム	0.872601	DBMS	0.963957
DBMS	0.725401	RDBMS	0.872601	データベース管理システム	0.963957
RDBMS	0.447808	DBMS	0.872601	ACID 特性	0.033629
リレーショナルデータベース管理システム	0.447808	データベース管理システム	0.872601	トランザクション処理	0.033629
階層型データベース	0.414156	折れ線グラフ	0.343659	原子性	0.033629
ネットワーク型データベース	0.414156	図解の種類と用途	0.343659	トランザクション管理	0.033629
データベース	0.367107	ガントチャート	0.343659	コミット	0.033629
データリポジトリ	0.358695	レーダーチャート	0.343659	融離性	0.033629
ボイス・コッド正規形	0.325517	Z グラフ	0.343659	アポート	0.033629
BCNF	0.325517	円グラフ	0.336717	定義域	0.033308

表 6　実験結果 A-2(コンテクスト：ジャーナル).

Table 6　Experimental results A-2 (Context:ジャーナル).

拡張空間		IT 分野の空間		DB 分野の空間	
ジャーナル	0.801441	ジャーナル	0.763818	ジャーナル	0.841556
ログ	0.530392	権利の保護	0.454610	ログ	0.841556
コミットコンシステントチェックポイント法	0.444281	知的所有権	0.454610	アンドゥ・リドゥ方式	0.344444
チェックポイント	0.444281	ソフトウェア	0.449352	アンドゥ	0.344444
アンドゥ	0.444281	ロールバック	0.420601	チェックポイント	0.279766
アンドゥ・リドゥ方式	0.444281	障害回復処理	0.420601	コミットコンシステントチェックポイント法	0.279766
ログファイル	0.357428	リカバリ	0.420601	キャッシュコンシステントチェックポイント法	0.054992
リカバリ	0.357428	ログファイル	0.420601	アクティブ	0.042547
障害回復処理	0.357428	システム環境の整備	0.275757	耐久性	0.042547
表の結合	0.340537	プレインストール	0.275757	B 木	0.032315

表 7　実験結果 A-3(コンテクスト：ファイル).

Table 7　Experimental results A-3 (Context:ファイル).

拡張空間		IT 分野の空間		DB 分野の空間	
ファイル	0.778414	OS	0.893266	ファイル	0.770400
OS	0.722069	絶対パス	0.870227	固定長レコード	0.770400
相対パス	0.697713	ファイル	0.870227	可変長レコード	0.770400
ディレクトリ	0.697713	相対パス	0.870227	レコード	0.481667
絶対パス	0.697713	フォルダ	0.870227	ブロック	0.410867
フォルダ	0.697713	ディレクトリ	0.870227	転送時間	0.410867
可変長レコード	0.692629	応用ソフトウェア	0.568543	バッファリング	0.410867
固定長レコード	0.692629	基本ソフトウェア	0.568543	ページ	0.410867
基本ソフトウェア	0.530764	ミドルウェア	0.568543	パックマン線図	0.296848
ミドルウェア	0.530764	ソフトウェア	0.282553	位置指示子	0.296848

「DBMS」と「データベース管理システム」のみ上位に出力しており，あとの語は相関量が非常に低い．これはデータベースの専門書であるため，データベース管理システムの説明が詳細に書いてある構成になっているために，構成上他の語と関連が低くなってしまったためと考えられる．拡張空間では,「データベース管理システム」と「DBMS」が同位 1 位「リレーショナルデータベース管理システム」と「RDBMS」が同位 3 位となり，IT 分野の空間で，この 4 語が同位 1 位になっていることと比べて，拡張により，関連がより詳細に表すことが可能になったと考えられる．さらに，IT 分野の空間，データベース関連の空間のどちらにも上位に出力されなかった，データベース管理システムと関連が深い,「データベース」「データリポジトリ」がそれぞれ，6 位，7 位に出力されている．拡張により，データ行列で関連付けされたためと考えられる．

コンテクスト「ジャーナル」の場合，表 6 から，IT 分野の空間では，書籍の構成上,「権利の保護」「知的所有権」が同位 2 位に出力されているが,「障害回復処理」「リカバリ」「ログファイル」などの関連のある語についても，同位 4 位に出力されている．DB 関連の空間では,「ログ」「アンドゥ・リドゥ方式」が 2 位，3 位に出力されている．拡張空間では，両者の出力のう

表 8　拡張空間による検索 1(コンテクスト：ファイル フォルダ).
Table 8　Experimental results by extended metadata space 1 (Context:ファイル フォルダ).

語	相関量
OS	0.788456
ディレクトリ	0.729396
相対パス	0.729396
フォルダ	0.729396
絶対パス	0.729396
ファイル	0.659789
基本ソフトウェア	0.583557
応用ソフトウェア	0.583557
ミドルウェア	0.583557
位置指示子	0.408711

表 9　拡張空間による検索 2(コンテクスト：ファイル レコード).
Table 9　Experimental results by extended metadata space 2 (Context:ファイル レコード).

語	相関量
固定長レコード	0.712737
可変長レコード	0.712737
ファイル	0.684923
レコード	0.557364
相対パス	0.532224
フォルダ	0.532224
ディレクトリ	0.532224
絶対パス	0.532224
レコード型	0.445396
位置指示子	0.445396

ち,「ジャーナル」に関連のある「ログ」「ログファイル」「リカバリ」「アンドゥ・リドゥ方式」「アンドゥ」「障害回復処理」が出力されていることがわかる.

コンテクスト「ファイル」の場合，表 7 から，IT 分野の空間では,「OS」「絶対パス」「相対パス」「フォルダ」「ディレクトリ」とファイルシステムに関連する用語が出力されている．DB 分野の空間では,「固定長レコード」「可変レコード」「レコード」「ブロック」と RDBMS が指すファイルに関連した語が出力されている．拡張空間では，その両方が組み合わせて出力している．これは両方のファイルの関連を表していると考えられる．

本環境は意味の数学モデルによる単語間関連連想検索を実現しているため，複数の語でコンテクストを表して，文脈を限定することにより，組み合わせて生成された拡張空間上で，個々の関係を表すことができると考えられる．

その例として，拡張空間での，コンテクスト「ファイル フォルダ」の検索結果を表 8，コンテクスト「ファイル レコード」の検索結果を表 9 に示す.

コンテクスト「ファイル フォルダ」は，表 8 より，ファイルシステムと関連する「OS」,「ディレクトリ」,「相対パス」.「絶対パス」が上位に出力されている.

一方，コンテクスト「ファイル レコード」は，表 9 より,「固定長レコード」,「可変長レコード」,「レコード」の順位がコンテクスト「ファイル」のときに比べて上がっている.

このことから，元の個々の空間の計量の機能をもちつつ，拡張していると考えられる.

5.4 実　験　B

5.4.1 実 験 方 法

IT 分野に関する画像メディアデータを対象として，単語間関連画像メディアデータ検索に適用し,「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】」[11] の索引部のみで生成したメタデータ空間と，上記と「データベースシステム」[12] の索引部で生成したデータ行列を組み合わせたメタデータ空間とで比較を行った．画像メディアデータは「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] に収録されている画像メディアデータ，418 個を対象とした．これらの画像メディアデータについて，手動でメタデータを付与した.

評価方法として，適合率という指標を用いて示した.

$$\text{適合率} = \frac{\text{システムの検索結果に含まれる正解数}}{\text{システムの検索結果出力数}} \tag{6}$$

なお，システムの検索結果出力数は上位 10 位とする.

5.4.2 実 験 結 果

10 のコンテクストについての，それぞれの結果を図 3 に示す．図中の「シスアドのみ」とは「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】」[11] のみで空間生成を行った場合,「シスアドと DB」は，これに「データベースシステム」[12] の二つで空間生成を行った場合である.

これらの結果から，データベースに関連しないコンテクストについては，同じ適合率となっている．それに対して，データベースに関連するコンテクストでは，適合率が同じか，高くなっている.

これにより，拡張することにより，拡張した分野の語をよりよく検索できることを示している.

5.5 考　　察

本実験から，各書籍の索引部を用いて生成したデータ行列を本方式を用いて統合することにより，扱うことの出来る語が増え，単語の関連をより詳細に計量可能なメタデータ空間となることを示した.

本実験では，IT 分野のデータ行列にデータベース関連の情報を拡張した．本拡張統合方式で，ネットワーク関連，ソフトウエア関連などの書籍の索引部を用いて拡張行うことによって，より詳細な関連の計量が可能なメタデータ空間となると考えられる.

6. あ と が き

本稿では，複数の書籍の索引部を用いたメタデータ空間生成拡張統合方式を示した．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メタデータ空間を生成することが可能となった．本方式を意味の数学モデルに適用することにより，語と語の関連を計量することによる，単語間の関連性に基づく連想検索である単語間関連連想検索を実現した.

本方式により，これまで，実現できなかった特定分野にも，連想検索の導入が容易に可能になると考えられる．さらに，特

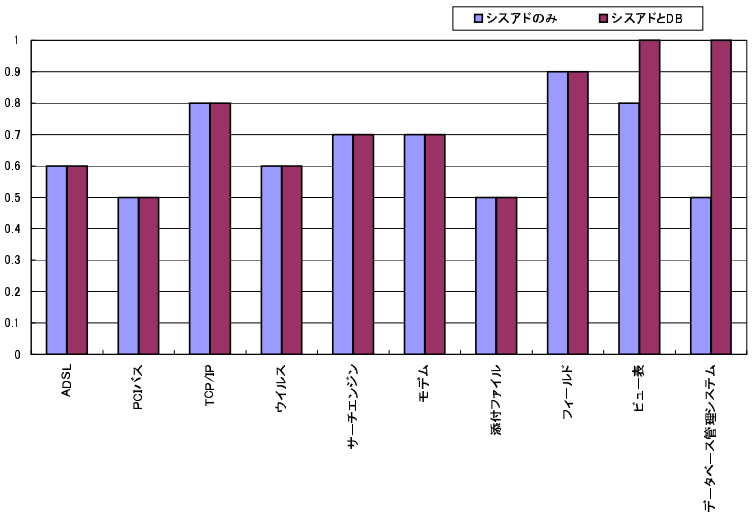

図 3　実験 B 結果.

Fig. 3　Experimental resultsB.

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，辞書や用語辞典が存在しない分野におけるメタデータ空間生成とその検索方式の実現，異種の情報源から生成されたメタデータ空間群の統合方式の実現が挙げられる．

文　　献

[1] Kitagawa, T. and Kiyoki, Y.: The mathematicalmodel 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).

[2] Kiyoki, Y., Kitagawa, T. and Hayama, T.: “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).

[3] 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).

[4] 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).

[5] 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).

[6] 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ページベースのメタデータ空間生成方式，データベースと Web 情報システムに関するシンポジウム (DBWeb2003),(2003).

[7] 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).

[8] Michael, W. B., Susan, T. D., Gavin, W. O.: Using linear algebra for intelligent information retrieval, SIAM Review Vol. 37, No.4, pp.573-595 (1995).

[9] Deerwester, S., Dumais, S., Furnas, G.W., Landauer, T.K. and Harshman, R.: Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science, Vol. 41, No. 6, pp.391-407 (1990).

[10] 伊東拓, 中西崇文, 北川高嗣, 清木康：“潜在的意味抽出方式と意味の数学モデルによる意味的連想検索方式の比較，” 第 13 回データ工学ワークショップ (DEWS2002) 論文集, 電子情報通信学会,(2002).

[11] 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).

[12] 北川 博之: “データベースシステム,” 情報系教科書シリーズ 第 14 巻 データベースシステム, (1996).

[13] “日経パソコン用語辞典 2004 CD-ROM 版, ” 日経 BP 社, (2003).