ドキュメント中に出現する単語間の関連性に基づく連想検索のためのメタデータ空間生成方式

本間　秀典† 　中西　崇文† 　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 情報数理研究室
†† 筑波大学大学院 システム情報工学研究科　〒 305–8577 茨城県つくば市天王台 1–1–1
E-mail: †{homma,takafumi}@mma.cs.tsukuba.ac.jp, ††takashi@cs.tsukuba.ac.jp

あらまし　ある特定の分野に関して意味の数学モデルによる連想検索を実現するためには，その分野を対象としたメタデータ空間と呼ばれる検索空間を生成する必要がある．これまで，メタデータ空間は，辞書や用語辞典，専門的な知識などを用いて生成していたが，対象とする分野に関する辞書や用語辞典が存在しない場合，その実現が困難であった．本稿では，ある特定分野を対象として，それらに関するドキュメントにおける単語の相対的な場所情報から計算される関連度によるメタデータ空間の自動生成方式について示す．本方式により，対象となる特定分野に関する単語間の関連を求めるメタデータ空間の自動生成が可能となる．本稿では，本方式を用いて生成したメタデータ空間に意味の数学モデルを適用し，その検索結果についても示す．

キーワード　メタデータ空間生成，検索空間，意味の数学モデル，単語間関連連想検索

A Construction Method of a Metadata Space for an associative search utilizing the relation of each word in documents

Hidenori HOMMA†, Takafumi NAKANISHI†, and Takashi KITAGAWA††

† Graduate School of Systems and Information Engineering, University of Tsukuba
†† Graduate School of Systems and Information Engineering, University of Tsukuba
E-mail: †{homma,takafumi}@mma.cs.tsukuba.ac.jp, ††takashi@cs.tsukuba.ac.jp

Abstract　In order to realize associative search for a specific field by the mathematical model of a meaning, it is necessary to establish the retrieval space called metadata space for the field. A metadata space was established using a dictionary, a term dictionary, and special knowledge. However, when neither a dictionary nor a term dictionary existed in the target field, it was difficult to establish metadata space. This paper presents a new construction method of a metadata space based on the relevance between words calculated from relative position information on them. This method enables establishment of a metadata space which measures the relation between words concerning a specific field. This paper shows the experimental results which applied the mathematical model of a meaning to the metadata space established using this method.

Key words　Establishment of a metadata space, Retrieval space, Mathmatical model of meaning, Words related associative search

1. ま え が き

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り，これらの情報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．

これまで，文献 [1]〜[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．データ行列の生成方式として，これまで文献 [2], [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．これらの方式では，辞書や用語辞典において説明される言葉（見出し語）の語義文中で見出し語の説明に使われている語（特徴語）を用いてその見出し語の特徴づけを行うことによって，意味を計量するためのメタデータ空間を生成し，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，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．

しかし，単語間の関連性の計量が可能なメタデータ空間生成ができれば，単語間の関連性に基づく連想検索，つまり，単語間関連連想検索が可能になると考えられる．

一般にドキュメントなどの文章では，読者が内容を理解しやすいように,「関係のある内容を近くにまとめて出現させることが多い」ということが，言語学などにおける語の意味に関する研究において言及されてきた [6]．ここで，これら「関係のある内容」は幾つかの文により表現され，それらの文は幾つかの単語の列によって構成されているので,「ドキュメント内においては関連性がある単語が近くにまとまって出現しやすい」と考えることができる．このような，ドキュメント中で出現する位置関係により単語の関連が現れる性質を用いてデータ行列を作成できれば，単語間の関連を計量する空間を容易に生成できると考えられる．しかも，ドキュメント内に現れる情報のみを利用してデータ行列の作成を行えば，辞書の作成のような高い専門性や多くの人手を必要とする作業を必要としないため，空間生成を自動化することができる．

本稿では, 単語同士の距離と頻度により計算される関連度によるメタデータ空間を生成する方式を示す．

本方式は，対象とする特定分野の教科書に相当するドキュメントを準備し，そのドキュメント内に出現する単語同士の関連性に注目してデータ行列を作成し，メタデータ空間を生成することを目的としている．これにより，辞書や用語辞典が存在しない分野において，語と語の関連性を表すメタデータ空間を自動的に生成できる．さらに，そのメタデータ空間を意味の数学モデル [1]〜[3] に適用することにより，単語間関連連想検索が実現できるため，文献 [2], [5] の方式の代替の検索方式として適用可能であると考えられる．

また，意味の数学モデルを用いた連想検索方式は，文献 [7], [8] 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 [9] で報告されている．

本稿では，出現する各単語の距離と頻度を用いたメタデータ空間生成方式について示す．さらに，本方式で生成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で，単語間関連連想検索を実現し，有効性の検証を行う．

2. ドキュメント中に出現する単語間の関連性に基づく連想検索のためのメタデータ空間生成方式

前節で述べたメタデータ空間を生成するためには，対象となる分野を反映したデータ行列を作成する必要がある．そこで本節では，特定分野に関するドキュメントから検索に利用する専門用語を自動抽出し，さらにそれらの間の相関を用いたデータ行列を自動生成するための方式について述べる．なお，ここでは対象となる分野に関するドキュメントが存在し，かつ何らかの方法で検索対象となるメディアからメタデータを抽出できることを前提としている．

2.1 ドキュメント中に出現する単語間の関連性

言語学者 Zellig Harris は，文章中の単語や形態素の意味について,「単語や形態素の意味の違いはそれらの分布の違いと相関がある」と述べている [6]. また，Harris は「分布的な性質は狭い範囲において見出せる」とも述べており，

これらの言及によれば，単語が出現する場合の距離と頻度から，それらの語と語の意味的な関連を知ることができると考えることができる．言い換えれば,「意味の近いものほどより物理的に近い場所に配置される頻度が高い」という傾向がある，と考えられる．これにより，ある概念を説明するために書かれたドキュメント内に出現するある語 w_1 とその近辺に出現する語 w_2 の関係について，次のような性質を考えることができる．

- ある単語 w_1 の近辺に出現する幾つかの単語がある場合，その出現位置が w_1 から近いものほど関連性が強い．
- ある単語 w_1 が同一ドキュメント内に複数回出現し，かつ w_1 から等しい距離に出現する単語が複数ある場合，出現する頻度の高いものほど関連性が強い．

単語間の関連に関するこれらの性質は，ドキュメント中に出現する各単語間の関連を求める上で重要であると考えられる．この性質を用いることによって対象となる分野に関するドキュメントから抽出した単語間の距離とその出現する頻度をもとにメタデータ空間を生成し，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できると考えられる．

以上の考察から，本方式では，特定分野に関する各単語間の関連性を抽出するために，その分野に関する任意のドキュメントを対象とすることができる．しかしながら，対象とするドキュメントにおいて，その分野に関して誤った記述がされていた場合，生成されるメタデータ空間の精度にも影響を及ぼしてしまうと考えられる．そのため，その分野の教科書に相当するような，誤った記述が無く信頼できるドキュメントを選択し対象とすることが重要であるが，ドキュメントの記述内容からその適切性を自動的に判断することは極めて難しい．そこで次節以降では，そのような「教科書に相当するドキュメント」を，人間の主観によって適切に選択することができたものとする．

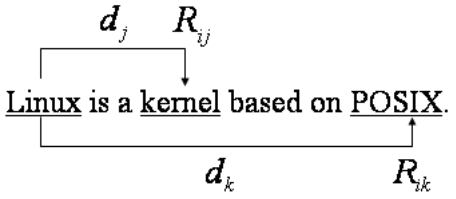

図 1　R_ij の概念
Fig. 1　An overview of R_ij

2.2　ドキュメント中に出現する単語間の関連性に基づくメタデータ空間生成

ここでは，対象となるドキュメント中に出現する単語間の関連性を用いたメタデータ空間生成方式を示す．その具体的な流れは以下のようである．

（1）　ドキュメントの解析

本方式では，対象となるドキュメントから検索語として用いる単語を抽出するために，専門用語自動抽出システム [11] を使用する．これにより，対象となるドキュメント中で重要であると思われる単語を自動的に抽出することができる．単語の抽出方式についての詳細は 2.3 節に述べられている．

（2）　単語間の関連性に基づく関連度の計算

次に，(1) で得られた N 語からなる語列に対し，ドキュメント中に出現する単語間の関連性に基づいて各単語の関連度を計算する．

はじめに，単語間の距離と，距離に基づく重みを設定する．まず，隣接して出現する 2 語間の距離を 1 とする．このとき，2 語が間に n 語を挟んで出現する場合の単語間の距離を d とすると，$d = n + 1$ となる．この d を用いて，2 語が隣接する場合を 1 とし，距離が大きくなるにつれて関連度が大きく下がるような評価関数 $W(d)$ を設定する．ここでは，予備実験により $W(d)$ を次式のように決定した．

$$W(d) = e^{1-d} \tag{1}$$

次に，以下の式により単語 w_j が単語 w_i から距離 d の位置に出現する頻度 $P_{ij}(d)$ を求める．

$$P_{ij}(d) = \frac{(w_j\text{が } w_i\text{から距離 } d \text{ に出現した回数})}{(w_i\text{がドキュメント内で出現した回数})} \tag{2}$$

式 (1)，(2) により，単語 w_i と w_j の関連を表す関数 R_{ij} は以下のように計算できる．

$$R_{ij} = \sum_{d=1}^{N-1} P_{ij}(d) \times W(d) \tag{3}$$

R_{ij} は w_i と w_j の出現頻度と出現時の距離に依存する関数であるから,「距離重み頻度関数」と呼ぶことにする．R_{ij} の概念は図 1 のようである．

ただし，

$$w_i w_i w_i \ldots w_i$$

のように，単一の単語 w_i のみからなる N 語の語列を考えると，$d = 1, 2, \ldots, N$ に対して $P_{ij}(d) = 1$ は明らかであり，しかも 3 つ以上同じ語が連続しても意味があるとは考えにくい．このことから，w_i と w_i の距離重み頻度関数値は以下に与えられるものとする．

$$R_{ii} = \sum_{d=1}^{3} W(d) \tag{4}$$

これにより，N 語の語列から重複して出現する単語を除いた語数を n とすると，式 (3)，(4) を用いて単語 w_i を特徴付けることができる．

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

以上から，$\mathbf{w}_i$ を用いて $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n)^T$ とすることによって，図 2 のような n 次正方行列 M を作成する．

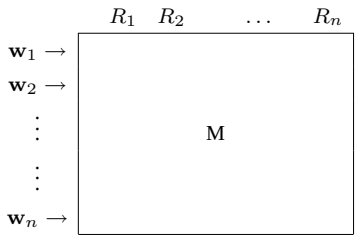

図 2　データ行列 M の表現
Fig. 2　An expression of a data matrix M

（3）　相関行列 M^TM からメタデータ空間生成

(2) で生成されたデータ行列 M の相関行列 M^TM を計算すると n 行 n 列の行列となる．この相関行列 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メタデータ空間を生成する．これにより，語と語の関係を計量する単語間関連連想検索のためのメタデータ空間の構成が可能となる．

2.3　ドキュメント中に出現する単語の抽出

特定分野に関するドキュメントから単語間の関連を抽出するためには，対象とする分野に関連する単語を抽出する必要がある．

ドキュメントから語を自動的に抽出するための方法として，形態素解析器を用いる方法が考えられるが，その分野に特有の専門用語が含まれるドキュメントを対象とする場合，適切な単語抽出が行えないということがしばしば起こり問題となる．例えば,「マルチタスク」「Web サーバ」「パーティション」という語を日本語形態素解析器「茶筌」[10] を用いて解析した場合,「マルチタスク」はそのまま 1 語として出力されるが,「Web サーバ」は「Web」と「サーバ」の 2 語に分解され,「パーティション」では「パーティ」「ション」と全く無関係か，もしくは意味を成さない文字素片に分解されてしまった．これらの語は IT 関連の話題における専門用語に相当するため，どの語も 1 語として抽出されるべき語である．この例から，専門用語を含む分野に関するドキュメントから単語を適切かつ自動的に抽出するためには，一般的な語を対象とした形態素解析では不十分で

あるということが分かる．ここで，対象となる分野に出現する語に関する辞書を追加して解析を行えればこの問題を回避できる可能性があるが，そのような辞書や用語辞典が存在しないなど，何らかの理由でそれを利用できない場合，単語の自動抽出を適切に行うことができない．

そこで，本方式では，専門分野に関するテキストから専門用語を自動抽出することを目的として東京大学中川研究室・横浜国立大学森研究室で共同開発されたシステムである,「専門用語自動抽出システム」[11] を利用する．このシステムは，対象となるテキストの形態素解析結果を用いて，独自の抽出理論によりドキュメント中における複合語の重要度を計算し，その重要度の高い複合語を専門用語として順に出力するものである．このシステムを用いた場合，茶筌のみによる解析では意味不明な文字素片まで分割されてしまった「パーティション」という語を 1 語として解析することができた．また，2 語に分割された「Web サーバ」，および適切に 1 語として解析できた「マルチタスク」という語についてもそれぞれ同様に 1 語として解析することができた．このことから，専門用語自動抽出システムは，特定分野の専門用語をドキュメントから自動的に抽出する方法として有効であることが分かる．

ここで，このシステムが，形態素解析の結果のみを要求しており，特定分野に関する辞書や何らかの語彙情報を必要としていない点にも注目したい．このシステムを用いることによって，一般の形態素解析のみでは用語の抽出を適切に行えない分野に関しても，辞書や用語辞典を用いることなく専門用語を自動的に抽出することができる．

3. 意味の数学モデルへの適用

本節では，2. 節で生成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]〜[3] に示している．

（1） 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．

$$D = \{\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_t\}. \tag{6}$$

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．

$$\mathbf{o}_i = (o_{i1}, o_{i2}, \cdots, o_{in}) \tag{7}$$

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．

$$\begin{aligned}\mathbf{d} = \bigoplus_{i=1}^{t} \mathbf{o}_i \\ := (\mathrm{sign}(o_{\ell_1 1}) \max_{1 \leqq i \leqq t} |o_{i1}|, \\ \mathrm{sign}(o_{\ell_2 2}) \max_{1 \leqq i \leqq t} |o_{i2}|, \\ \cdots, \mathrm{sign}(o_{\ell_n n}) \max_{1 \leqq i \leqq t} |o_{in}|).\end{aligned} \tag{8}$$

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．

$$\max_{1 \leqq i \leqq t} |o_{ik}| = |o_{\ell_k k}|. \tag{9}$$

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．

（2） 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．

4. 実　　験

提案方式の有効性を示すため，提案方式に基づくデータ行列生成システムを構築し実験を行った．

4.1 実験環境

本実験において，専門用語自動抽出システムを利用するために，日本語ドキュメントの全文検索などにおいて広く用いられている形態素解析器 Chasen [10] を用いた．また，メタデータ空間を生成するためのドキュメント群としては，Web サイト「@IT」[12] より Linux 関連の連載 11 記事 22 ページを使用した．

4.2 実　　験 1

4.2.1 実験方法

実験 1 では，提案方式によって生成されるメタデータ空間の性質を，他の方式によって生成したメタデータ空間と比較してその性質を確認する．提案方式を前出の Web ページ [12] の記事に適用してメタデータ空間を生成し，検索語，検索対象とも同ページ中に出現する専門用語として検索実験を行った．さらに，提案方式との比較を行うために，ページごとに出現する単語の TF-IDF 値を用いてデータ行列の特徴付けを行うことによってメタデータ空間を生成し，提案方式を用いた場合と同じ検索語を用いて検索実験を行った．後者の方法では，ページ $\mathbf{p}_i$ を出現する n 個の単語 $w_k (k = 1, 2, \cdots, n)$ の TF-IDF 値 V_k を用い

表 1　実験結果 1-1（コンテクスト：Linux パーティション）
Table 1　Experimental results 1-1 (Context: Linux パーティション)

順位	語	相関量
1	パーティション	1.000000
2	Linux	1.000000
3	(以下，検索結果無し)	0.000000

表 2　実験結果 1-2（コンテクスト：Linux パーティション）
Table 2　Experimental results 1-2(Context: Linux パーティション)

順位	語	相関量
1	Linux	0.469070
2	パーティション	0.423599
3	HDD	0.347568
4	Mbytes	0.329853
5	リスト	0.329557
6	UNIX	0.309279
7	hosts	0.296713
8	アクセス制御	0.281672
9	bash	0.274675
10	ファイルシステム	0.264867

て次のように特徴付ける．

$$\mathbf{p}_i = (V_1, V_2, \cdots, V_n) \tag{10}$$

この $\mathbf{p}_i$ を用いて $(\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_n)^T$ とすることによって，図 1 と同様なデータ行列 M を作成する．この行列 M を 3. 節で示した方法に適用することにより，提案方式との比較を行う．

4.2.2　実 験 結 果

表 1，2 はそれぞれ TF-IDF，および提案方式を用いてメタデータ空間を生成して検索実験を行った結果の上位 10 件を示している．このとき，どちらの実験においても，「Linux パーティション」の 2 語を検索語としている．ただし，表 1 においては，3 番目以降の検索結果の相関量が 0 であったため省略されている．

4.2.3　考　　察

表 1 によると，TF-IDF 値によるデータ行列の特徴付けを行った場合，検索語のみが検索されている．これは，対象としたドキュメント数が少ないためと考えられる．そのため，対象とするドキュメント数を大きくすることによって結果を改善できる可能性があるが，単語間の関連を適切に反映するのは難しいと考えられる．

これに対し，表 2 によると，提案方式では，対象としたドキュメントの内容と合致する，検索語と関連の強い語が検索されていることが分かる．この実験は，提案方式を用いることにより，ドキュメントの内容を反映したメタデータ空間を生成できることを示している．

4.3　実　験　2

4.3.1　実 験 方 法

実験 2 では，提案方式を用いてメタデータ空間を生成した場合の検索精度の検証を行う．

まず，提案方式を用いて実験 1 同様にメタデータ空間を生成する．このとき，対象となる分野に関する擬似データを手動で作成し，それらを検索対象とする．ここでは，以下のように，擬似データは対象となる分野に属する 10 のカテゴリ 0 〜 9 の中からそれぞれ 5 件ずつ，計 50 件作成し，それぞれにメタデータとして単語を 3 語ずつ設定した．

- ID 00 〜 04 ：　カテゴリ 0 に属する擬似データ
- ID 10 〜 14 ：　カテゴリ 1 に属する擬似データ
- ID 20 〜 24 ：　カテゴリ 2 に属する擬似データ
- ⋮
- ID 90 〜 94 ：　カテゴリ 9 に属する擬似データ

また，検索語はそれぞれのカテゴリに関連する 2 語とし，そのカテゴリに属する 5 件のデータをその検索語に対する検索結果の正解とした．ここで設定した擬似データの例を表 3 に示す．

以上の条件で検索実験を行い，検索結果から再現率，適合率を計算することにより検索精度の検証を行った．

表 4　実験 2 の検索語
Table 4　Keyword for Experiment 2

実験番号	検索語	対応カテゴリ
2-1	シェルスクリプト　制御	9
2-2	パターンマッチ　文字列	8
2-3	コマンドライン　エディタ	7

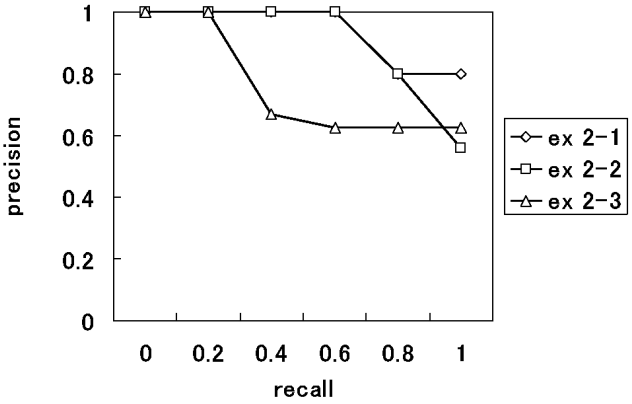

図 3　実験結果 2（再現率-適合率グラフ）
Fig. 3　Experimental result 2 (recall-precision graph)

4.3.2　実 験 結 果

実験結果から作成した再現率-適合率グラフを図 3 に示す．図 3 の実験 2-1，2-2，2-3 それぞれにおいて検索実験に用いた検索語，およびそれぞれに対応するカテゴリは表 4 に示すとおりである．

4.3.3　考　　察

図 3 から，提案方式を用いることにより，どの再現率レベルにおいても高い適合率が維持できていることが分かる．この結果は，提案方式がドキュメント中に出現する単語間の関連を適切に反映したメタデータ空間を生成し，単語間の関連に基づく連想検索を実現できることを示している．

表 3　実験 2 で作成した擬似データの例
Table 3　Imputed Data for Experimet 2

データ ID	メタデータ
00	POSIX　カーネル　API
01	OS　リソース 管理
02	OS　中核部分　処理
03	カーネル　プログラム　追加
04	POSIX　互換性　API
10	マルチユーザー　コンポーネント指向　ユーザーインターフェイス
11	堅牢性　遠隔地　CUI
12	マルチタスク　マルチユーザー　システム管理
13	ユーザー　仮想端末　複数
14	システム管理　メンテナンスコスト　ログイン
20	Linux　ファイルシステム　管理
21	Windows　ドライブ　fdisk
22	ファイル名空間　ディレクトリ構造　ファイルシステム
23	バックアップ　アップデート　ユーザー
24	パーティション　使用率　偏り
⋮	⋮
90	繰り返し　実行　処理
91	シェルスクリプト　実行属性　chmod
92	制御　処理対象　引数
93	test コマンド　シェルスクリプト　条件式
94	ループ　省力化　制御構造

4.4　実験のまとめ

実験 1 では，提案方式により生成されるメタデータ空間を用いて検索実験を行い，提案方式によりドキュメント中に出現する単語間の関連性を反映した検索結果が得られることを示した．さらに，他の方式との比較を行うことにより提案方式の有効性を示した．

実験 2 では，再現率-適合率グラフにより提案方式を用いてメタデータ空間を生成した場合の検索精度の検証を行った．これにより，提案方式が高い検索精度を実現できることを示した．しかしながら，この実験は，ある特定分野の一部に関するドキュメントのみを利用した場合の結果に過ぎない．そのため，特定分野全体を包含したドキュメント群を用いた検索実験，およびその検索精度の検証は今後の課題である．

5.　あ と が き

本稿では，ドキュメント中に出現する単語間の関連性に基づくメタデータ空間生成方式を示した．また，提案方式をドキュメント群に適用して検索実験を行うことにより，その検索精度の検証を行った．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，対象とする特定分野の教科書に相当するドキュメントを準備し，そのドキュメント中に出現する単語同士の関連性に注目してデータ行列を作成し，メタデータ空間を生成することが可能となった．これにより，辞書や用語辞典が存在しない分野において，語と語の関連性を表すメタデータ空間を自動的に生成できる．

今後の課題として，ある特定分野全体を包含したドキュメント群への提案方式の適用，および，より多数のデータを検索実験に用いた，より信頼できる規模での検索精度の検証，辞書や用語辞典を用いてメタデータ空間を生成した場合など既存の方式との比較などが挙げられる．

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(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] Longman Dictionary of Contemporary English, Longman (1987).

[5] 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).

[6] Zellig S Harris / edited by Henry Hiz. Reidel：“Papers on syntax” (1981)

[7] 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).

[8] 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).

[9] 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).
[10] http://chasen.naist.jp/
[11] http://gensen.dl.itc.u-tokyo.ac.jp/index.html
[12] http://www.atmarkit.co.jp