DEWS2004 2-B-04

ドキュメント内における単語の局所性を用いた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 第三学群 情報学類　〒 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 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 電子・情報工学系　〒 305–8577 茨城県つくば市天王台 1–1–1
E-mail: †{homma,takafumi}@nalab.is.tsukuba.ac.jp, ††takashi@is.tsukuba.ac.jp

あらまし　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，専門的な知識を用いて生成していた．しかしながら，対象とする分野に辞書や用語辞典が存在しない場合，メタデータ空間を生成することが困難であった．本稿では, 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を生成する方法を示す．本方式を用いることにより，単語間の関連を求めるメタデータ空間を，辞書を作成することなく容易に生成できる．その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，単語間の関連性に基づく連想検索である，単語間関連連想検索が実現できる．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，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 locality of words in documents

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

† College of Information Sciences, Third Cluster of Colleges, University of Tsukuba
†† Graduate School of Systems and Information Engineering, University of Tsukuba
††† Institute of Information Sciences and Electronics, University of Tsukuba
E-mail: †{homma,takafumi}@nalab.is.tsukuba.ac.jp, ††takashi@is.tsukuba.ac.jp

Abstract　In order to realize associative search for a specific field by the mathematics model of a meaning, it is necessary to establish the retrieval space called metadata space for the specific 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 locality of words in documents. This method enables establishment of a metadata space which measures the relation between words easily without making any dictionary. The words related associative search for documents and mediadata of a specific field is realized by applying the metadata space to the mathematics model of a meaning. This paper shows the experimental results which applied the mathematics 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る．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．

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．これらの方式によって，意味を計量するための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，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．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，場所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[6],[7]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[8]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. ドキュメント内における「単語の局所性」を用いた連想検索のためのメタデータ空間生成方式

本節では，ドキュメント内に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の提案方式を示す．本方式では，検索対象が包含する特定分野，およびその分野について書かれたドキュメントが存在することを前提としている．

2.1節では，ドキュメント内における単語の出現傾向であると考えられる,「単語の局所性」について考察する．2.2節では，2.1節を受けて，ドキュメント内における「単語の局所性」を用いたメタデータ空間生成方式の実現について示す．

2.1 「単語の局所性」

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 について，次の場合が考えられる．

- w_1 を表現するために w_2 が用いられている．
- w_1 が w_2 を表現するために用いられている．
- w_1，w_2 を用いてある概念 P が表現されている．

以上のどの場合においても，ある語 w_1 とその近辺に出現する語 w_2 はある一つの概念を表現するために用いられており，何らかの関連性があると考えられる．このように，ある概念について書かれたドキュメントでは，読者が内容を理解し易いように，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．ドキュメント内に出現する単語に関しても同様に，ある内容を表現するために関連性がある幾つかの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，ある語が出現することによって何らかの関連性を持つ幾つかの語がドキュメント内で局所的に出現する性質を「単語の局所性」と呼ぶことにする．また，何らかの関連性を持つ語句が近くに集まり易いことから，似た内容を包含する概念を説明する際には同じ単語を繰り返し用いる可能性が高いと考えられる．

以上の考察から，以下の２つの性質が言える．

- 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，出現する確率の高いものほど関連性が強い．

このことから，単語の局所性はドキュメント内におけ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．

これにより，対象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確率をもとにメタデータ空間を生成できれば，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．

次節から，ドキュメント内における単語の局所性を用いてメタデータ空間を自動生成する方式を示す．

2.2 単語の局所性を用いたメタデータ空間生成

ここでは，対象となるドキュメント内における単語の局所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

まず，対象となるドキュメントに対して形態素解析を行う．本方式では，日本語ドキュメントの検索などの研究において形態素解析に広く用いられている ChaSen [9] を利用する．これにより，対象となるドキュメントからその品詞情報を付加された出現単語の列が得られる．この品詞情報を参照することによって，助詞や助動詞など検索を行う上であまり意味を表さないと考えられる単語を除去することができる．

この処理によって得られた N 語の語列に対し，さらに以下の処理によって各単語間の関連度を計算する．

（2） 単語の局所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，単語の局所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 n とすると，$d = n + 1$ となる．この d を用いて，単語間の距離に応じた関連度を評価する．隣接する場合の関連度を 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に，以下のように評価関数 $W(d)$ を設定する．

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

なお，評価関数 $W(d)$ については，予備実験によって幾つかの候補から最も結果が良かったものを選んだ．しかしながら，他の関数も適用可能であると考えられるため，この評価関数 $W(d)$ による評価は今後の課題である．

次に，以下の式により単語 w_j が単語 w_i から距離 d の位置に出現する確率 $P_{ij}(d)$ 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 の関連度 R_{ij} は以下のように計算できる．

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

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 の関連度は以下に与えられるものとする．

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

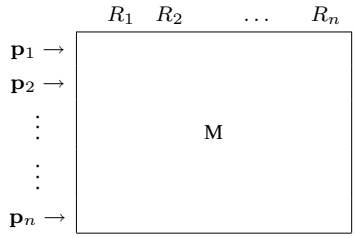

図 1 データ行列 M によるメタデータの表現

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．

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

以上から，$\mathbf{p}_i$ を用いて $(\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n)^T$ 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．

（3） 行列 M からメタデータ空間生成

(2) で生成されたデータ行列 M は単語と単語の関係を示す行列となる．これを固有値分解して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．

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．

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

$$\cdots, \mathrm{sign}(o_{\ell_n n}) \max_{1 \leqq i \leqq t} |o_{in}|). \tag{8}$$

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

本提案方式の有効性を検証するため，2. 節で示した方式を用いて生成したメタデータ空間について検証実験を行った．

本提案方式は，専門家によって人手で作成された辞書のような適切な辞書や用語辞典などが存在しない分野において単語間関連連想検索を行うためのメタデータ空間の生成を想定している．そこで，本実験では，適切な辞書が存在しない分野を対象としたメタデータ空間生成を行うことによって，提案方式の検証と性質の考察を目的としている．

実験１では，同じような構造を持つ２つの文章からなる短い文章に提案方式を用いて生成したメタデータ空間による単語間関連連想検索を行い，本方式により生成されるメタデータ空間の性質の検証を行った．

実験２では，計画問題を解くための数学の分野である数理計画法に関する複数のドキュメントから本方式によりメタデータ空間を生成し，数式を検索対象とした単語間関連連想検索を実現し，本方式によるメタデータ空間の性質，及び有効性の検証を行った．

4.1 実 験 環 境

それぞれの実験において，HTML 形式のドキュメントを対象にメタデータ空間の生成，および検索実験を行った．ここで，ドキュメントから単語の抽出を行うための形態素解析器には，2.2 節で述べたように ChaSen を用いた．解析時には ChaSen の標準的な辞書を用いたため，実験 2 では解析時に元通りに抽出できずに分解されてしまった専門用語も幾つか存在した．

自己紹介

私の名前は本間秀典です.
私の趣味はテニスです.
私の好きな動物は犬です.

彼の名前は山田太郎です.
彼の趣味は将棋です.
彼は動物は嫌いです.

図 2　実験に用いた文章
Fig. 2　Document used for this experiment

実験 1 では，図 2 に示すような文章を作成して実験を行った．提案方式によって 12 次元のメタデータ空間が生成された．

実験 2 では，数学の分野のひとつである数理計画法に関する Web ページ「数理計画のオンライン・テキスト」[10] を利用して実験を行った．提案方式によって 2637 次元のメタデータ空間が生成された．さらに，同ページ内に出現する数式のうちの 101 個を検索対象として，その数式に対して手動でメタデータを付与した．検索対象の数式とそのメタデータの一部を表 4 に示す．

4.2 実　験　1

図 2 に示す文章を用いて提案方式により生成されるメタデータ空間での単語間関連連想検索により，その性質を検証した．

4.2.1 実 験 方 法

提案方式である単語の局所性を用いたメタデータ空間を生成し，さらにそれを用いてドキュメント内に出現する幾つかの表現をコンテクストとして単語間関連連想検索を行った．出現する単語をそのまま検索結果として用いることにより，与えられたコンテクストと関連のある単語が適切に出力されるかどうかを検証した．なお，ここで利用した文章は単純な構造になっており，検索結果が正解といえるかどうかは明らかである．

4.2.2 実 験 結 果

図 2 に示した文章から本提案方式により生成したメタデータ空間による単語間関連連想検索の結果として，コンテクスト「名前」「彼 名前」「私 動物」の 3 つの場合を，それぞれ表 1, 2, 3 に示す．

コンテクスト「名前」の場合，表 1 から,「名前」という単語と関連があると思われる,「秀典」,「本間」,「太郎」,「山田」といった語が結果の上位に来ている．次に，コンテクスト「彼 名前」の場合，表 2 から,「太郎」，そして「山田」が 1，2 位に来ている．この２つの場合の検索結果から，図 2 のドキュメント内において,「名前」は「本間秀典」か「山田太郎」のどちらかであり，そのうち「彼の名前」は「山田太郎」である，という単語間の関連を適切に表現できているといえる．

最後に，表 3 はコンテクスト「私 動物」の場合の検索結果である．同表によれば，最上位には「好き」「犬」の 2 つが現れている．これは文中における「私の好きな動物は犬」であるという記述を反映したものであり，本文における単語間の関連を連

表 1　実験結果 1-1（コンテクスト：名前）
Table 1　Experimental results 1-1 (Context:名前)

順位	相関量	語句
1	0.876845	名前
2	0.871504	秀典
3	0.871504	本間
4	0.726638	太郎
5	0.726638	山田
6	0.627858	彼
7	0.421974	私
8	0.281669	将棋

表 2　実験結果 1-2（コンテクスト：彼 名前）
Table 2　Experimental results 1-2(Context:彼 名前)

順位	相関量	語句
1	0.815053	太郎
2	0.815053	山田
3	0.712537	彼
4	0.592154	名前
5	0.309694	将棋
6	0.102127	嫌い
7	0.099157	動物
8	0.077483	犬

表 3　実験結果 1-3（コンテクスト：私 動物）
Table 3　Experimental results 1-3 (Context:私 動物)

順位	相関量	語句
1	0.798218	好き
2	0.798218	犬
3	0.781812	私
4	0.559800	動物
5	0.255368	テニス
6	0.103778	嫌い
7	0.071817	彼
8	0.069282	秀典

想できた結果が現れていると考えることができる．

4.2.3　考　　察

本実験によって，本提案方式が，ドキュメント内に出現する単語間の関連を反映したメタデータ空間を生成し，それにより単語間関連連想検索を実現できるということが示された．本実験の結果から，本提案方式によって，検索語列と最も関連の強い語句を，ドキュメント内に出現する単語間の関連に基づいて検索することができたと考えられる．

4.3　実　験　2

「数理計画のオンライン・テキスト」[10] を利用してメタデータ空間を生成し，同ドキュメント中に現れる数式に手動でメタデータを割り当てたものを検索対象として検索実験，および検証を行った．検索対象とした数式の一例を図 4 に示す．

4.3.1　実 験 方 法

本提案方式により生成されたメタデータ空間と，それを用いて手動で設定した検索対象の数式を用いて検索実験を行った．

また，検索に用いる各コンテクストについて，設定した数式のうち幾つかを正解とし，これを検索結果の評価の基準とした．正解の設定は Web ページ内での数式の説明を参考に，以下の判断により手動で行った．

コンテクスト「双対」は，計画問題の双対問題や双対性，双対定理といった内容を連想するコンテクストである．検索対象に選んだ式から，これらに当てはまると思われる 10 個を正解とした．

コンテクスト「基底 選択」は，基底解を求めるために基底変数，もしくは非基底変数を選択する方法に関連したコンテクストである．ここでは，基底変数やピボットの選択規則，及びそれにより導かれると考えられる 20 個の式を正解とした．

コンテクスト「支持 平面」は,「支持超平面」を想定したコンテクストである．文中に「凸関数 $f(\boldsymbol{x})$ の共役関数 $f^*(\boldsymbol{u})$ とは，傾斜$\boldsymbol{u}$の超平面（線形関数　$\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}-\alpha$）によって $f(\boldsymbol{x})$ を下から支える 支持超平面を表している」とあることから,「支持超平面」，および「共役関数」に関連する 2 つの式を正解とした．

さらに，本実験によって，本方式により生成されたメタデータ空間の性質，及び専門的な辞書などを利用できない分野における本方式の有効性を検証するために，検索結果の考察を行った．

4.3.2　実 験 結 果

本方式により生成したメタデータ空間による単語間関連連想検索の検索実験の結果として，コンテクスト「双対」,「基底 選択」,「支持 平面」の 3 つの場合をそれぞれ表 5，6，7 に示す．これらの表はそれぞれの場合において検索結果の上位 10 位を示している．

コンテクスト「双対」の場合，表 5 からも分かるように，上位の多くに適合する式が検索されている．それら適合する式の殆どがメタデータにコンテクストと同じ「双対」を含んでいることからパターンマッチングに近い結果が得られたと考えられるが，9 番目に検索された式は本文にある「主問題の最適解は共役関数を用いた計画問題を解いても得ることができ，この問題を双対問題という.」という文脈から連想されていると考えられ，単語間の関連も連想できていると推察できる．

コンテクスト「基底 選択」の場合，表 6 によると，シンプレックス法におけるピボット選択規則に関する 2 つの式が検索結果の上位 1，2 位に来ている．3 位，及び 5 位は基底変数の選択とはあまり関係の無い式だが，残りの 4 位，及び 6 位から 10 位の式は全て計画問題の標準形式から基底解を求めるための方法に関する式である．特に，4，6，7，9 位の式については，メタデータとして与えた語にコンテクストの 2 語のうちのいずれも含んでいないにも関わらず上位に来ていることから，単語間の関連を適切に連想できていると考えられる．

コンテクスト「支持 平面」の場合，表 7 からも分かるように，正解がまったく得られなかった．これは,「平面」という語がドキュメント中でかなり頻繁に出現し，様々な語と関連性を持っているため,「支持」という単語との関連を的確に連想できなかったためであると考えられる．しかしながら，形態素解析で「支持超平面」という語句を一語として扱えなかったことも大きな要因の一つであったと考えられるので，まだ改善の余地

表 4　検索対象の数式とメタデータの一部

Table 4　Expressions for search and a part of metadata

検索対象の数式	メタデータ
$\exists x \in X \mid x = \lambda_{x_1} + (1-\lambda)_{x_2}; \forall x_1, \forall x_2 \in X, 0 \leqq \forall\lambda \leqq 1$	線形 凸 任意
$x_1, x_2 \in X \mid x = \lambda_{x_1} + (1-\lambda)_{x_2}; x \in X, 0 < \lambda < 1$	端点 境界 孤立
$c_D^t = c_N^t - c_B^t D$	相対 費用 係数
$a'_{iq} = 0, a'_{ij} = a_{ij} - \frac{a_{iq} * a_{pj}}{a_{pq}}, (i \neq p, j \neq q)$	枢軸 変換 演算
$\pi^t = c_B^t B^{-1}$	シン プレックス 乗数

表 5　実験結果 2-1（コンテクスト：双対）

数式	相関量	メタデータ	正解
$-\infty \leqq \phi(\boldsymbol{x}^*, 0) = \psi(0, \boldsymbol{v}^*) \leqq \infty$	0.315969	強い 双対 定理	○
$\inf_{\boldsymbol{x}}\{\phi(\boldsymbol{x}, 0) \mid \boldsymbol{x} \in \mathbf{R}^n\} \geqq \sup_{\boldsymbol{v}}\{\psi(0, \boldsymbol{v}) \mid \boldsymbol{v} \in \mathbf{R}^m$	0.300345	双対 摂動 共役	○
$\boldsymbol{b}_{B,p} = \min(k; \boldsymbol{b}_{B,k} < 0)$	0.294099	双対 シン 枢軸	○
$\mathbf{P}^* : \max_{\boldsymbol{v}} \psi(0, \boldsymbol{v}), \boldsymbol{v} \in \mathbf{R}^m$	0.292127	双対 共役 関数	○
$\boldsymbol{c}^{\mathrm{t}} - \boldsymbol{\pi}^{\mathrm{t}} A \geqq 0$	0.264718	標準 主 双対	○
$\mathbf{P}^*(\boldsymbol{u}) : \max_{\boldsymbol{v}} \psi(\boldsymbol{u}, \boldsymbol{v}), \boldsymbol{u} \in \mathbf{R}^n, \boldsymbol{v} \in \mathbf{R}^m$	0.259003	摂動 双対 問題	○
$\underline{f}(\boldsymbol{x}, 0) = f(\boldsymbol{x}), \boldsymbol{x} \in \mathbf{R}^n$	0.240910	摂動 関数 凸計画	×
$\frac{\boldsymbol{c}_{D,s}}{d_{ps}} = \max(\frac{\boldsymbol{c}_{D,l}}{d_{pl}}; d_{pl} < 0)$	0.235904	双対 シン 変数	○
$f^*(\boldsymbol{u}) = \inf_{\alpha}\{ \mid \boldsymbol{u}^{\mathrm{T}}\boldsymbol{x} - f(\boldsymbol{x}) \leqq \alpha, \forall \boldsymbol{x} \in \mathbf{R}^n, \alpha \in \mathbf{R}^1\}, \boldsymbol{u} \in \mathbf{R}^n = \sup_{\boldsymbol{x}}\{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x} - f(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbf{R}^n\}, \boldsymbol{u} \in \mathbf{R}^n$	0.228077	共役 関数 支持	○
$\phi(\boldsymbol{x}, \boldsymbol{y}) = \{ \begin{matrix} \underline{f}(\boldsymbol{x}, \boldsymbol{y}) \mid \underline{g}(\boldsymbol{x}, \boldsymbol{y}) \leqq 0 \\ \infty \mid \text{その他の場合} \end{matrix}$	0.216682	摂動 関数 拡張	×

表 6　実験結果 2-2（コンテクスト：基底 選択）

数式	相関量	メタデータ	正解
$\boldsymbol{c}_{D,s} = \min(l; \boldsymbol{c}_{D,l} < 0)$	0.348189	シン 枢軸 選択	○
$\frac{\boldsymbol{b}_{B,p}}{d_{ps}} = \min(\frac{\boldsymbol{b}_{B,k}}{d_{ks}}; d_{ks} > 0)$	0.330454	式 番号 選択	○
$\begin{pmatrix} a & b \\ d & e \end{pmatrix} (\boldsymbol{x}, -1)^{\mathrm{t}} = \left\{ \begin{matrix} O \\ 1 \end{matrix} \right\} f$	0.204788	拡大 係数 行列	×
$A = [B, B_N], A \in \mathbf{R}^{m*n}, B \in \mathbf{R}^{m*n}, B_N \in \mathbf{R}^{m*(n-m)}$	0.200263	係数 分割 行列	○
$\exists \boldsymbol{x} \in \mathbf{X} \mid \boldsymbol{x} = \lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2 \ ; \ \forall \boldsymbol{x}_1, \forall \boldsymbol{x}_2 \in \mathbf{X}, 0 \leqq \forall_\lambda \leqq 1$	0.198745	線形 凸 任意	×
$\boldsymbol{c}^{\mathrm{T}} = (\boldsymbol{c}_B^{\mathrm{T}}, \boldsymbol{c}_N^{\mathrm{T}}), \boldsymbol{c} \in \mathbf{R}^n, \boldsymbol{c}_B \in \mathbf{R}^m, \boldsymbol{c}_N \in \mathbf{R}^(n-m)$	0.198308	係数 分割 ベクトル	○
$B_{\boldsymbol{X}_B} + B_N \boldsymbol{X}_N = \boldsymbol{b}$	0.193583	ベクトル 行列 標準	○
$\boldsymbol{x}_B = (\boldsymbol{x}_{i1}, \cdots, \boldsymbol{x}_{im})^{\mathrm{T}}, I_B = (i_1, \cdots, i_k, \cdots, i_m), k = 1, \cdots, m$	0.190191	標準 基底 変数	○
$\boldsymbol{b}_B = B^{-1}\boldsymbol{b}$	0.189014	標準 正 ベクトル	○
$f = f_0 + \boldsymbol{c}_D^{\mathrm{T}} \boldsymbol{x}_N \rightarrow \min(\text{or } \max)$	0.186784	目的 非 基底	○

表 7　実験結果 2-3（コンテクスト：支持 平面）

数式	相関量	メタデータ	正解
$\exists \boldsymbol{x} \in \mathbf{X} \mid \boldsymbol{x} = \lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2; \forall \boldsymbol{x}_1, \forall \boldsymbol{x}_2 \in \mathbf{X}, 0 \leqq \forall\lambda \leqq 1$	0.388306	線形 凸 任意	×
$\boldsymbol{b}_B = B^{-1}\boldsymbol{b}$	0.353899	標準 正 ベクトル	×
$\boldsymbol{c}^{\mathrm{T}} = (\boldsymbol{c}_B^{\mathrm{T}}, \boldsymbol{c}_N^{\mathrm{T}}), \boldsymbol{c} \in R^n, \boldsymbol{c}_B \in \mathbf{R}^m, \boldsymbol{c}_N \in \mathbf{R}^{n-m}$	0.351952	係数 分割 ベクトル	×
$a_{11}\boldsymbol{x}_1 + a_{12}\boldsymbol{x}_2 + \cdots + a_{1n}\boldsymbol{x}_n - \boldsymbol{x}_{n+1} = \boldsymbol{b}_1$	0.319275	線形 余裕 変数	×
$B_{\boldsymbol{x}_B} + B_N \boldsymbol{x}_N = \boldsymbol{b}$	0.307004	ベクトル 行列 標準	×
$S_0 = (\boldsymbol{x} \mid g_i(\boldsymbol{x} < 0, \forall i \in \mathbf{I})$	0.304459	非線形 不等号 境界	×
$\boldsymbol{x}_B = (x_{i1}, \cdots, x_{im})^{\mathrm{T}}, \mathbf{I}_B = (i_1, \cdots, i_k, \cdots, i_m), k = 1, \cdots, m$	0.294962	標準 基底 変数	×
$f(\lambda\boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) \leqq \lambda f(\boldsymbol{x}_1) + (1-\lambda) f(\boldsymbol{x}_2)$	0.293463	凸 関数 下向き	×
$a_{11}\boldsymbol{x}_1 + a_{12}\boldsymbol{x}_2 + \cdots + a_{1n}\boldsymbol{x}_n + \boldsymbol{x}_{n+1} = \boldsymbol{b}_1$	0.284018	線形 不足 変数	×
$x_1, x_2, \cdots, x_n \geqq 0$	0.281735	線形 標準 非負	×

があるといえる．

4.3.3　考　　察

本実験により，提案方式によって生成されたメタデータ空間で単語間関連連想検索を行うことによってコンテクストと関連のある語をメタデータとして持つ数式を上位に出力させることができる，ということが示せた．

しかしながら，単語の局所性だけでなく，その内容まで考慮してコンテクストを発行すると，その内容を反映できない場合もあった．これは，単語間の距離によるため，近くにある語全てを関連があるとしてしまう性質が影響していると考えられる．また，単語の局所性だけでなく，その内容を考慮してメタデータ空間の生成を行わないと内容の不足を反映できない場合があるため，内容を考慮したメタデータ空間との連携は今後の課題であるが，本方式のみでも十分通用すると思われる．

4.4 実験全体のまとめ

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単語間関連連想検索を行うことによって性質を検証し，有効性を示した．

実験２では，メタデータに設定した単語とコンテクストの関連を連想して検索を行う場合の，提案方式によるメタデータ空間の性質を検証し，その有効性を示した．

この実験は，ドキュメント内における単語の局所性を用いたメタデータ空間生成方式の有効性を示している．

5. あ と が き

本稿では，ドキュメント内における単語の局所性を用いたメタデータ空間生成方式を示した．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，実際の特定分野の検索対象データを対象とした単語間関連連想検索の実現とその定性的な検索精度の検証，内容を考慮したメタデータ空間との連携が挙げられる．

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] 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] 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).

[7] 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).

[8] 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).

[9] http://chasen.aist-nara.ac.jp/

[10] http://lecture.ecc.u-tokyo.ac.jp/ õkatu/planning/01pl _ contents.html