2004－DBS－134　(II)　(67)
2004／7／14

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†††

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の生成が容易に可能となる．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．

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†††

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.

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

† 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：homma@nalab.is.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
Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan
e-mail：takafumi@nalab.is.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
Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan
e-mail：takashi@is.tsukuba.ac.jp

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．

ここで，一般的なドキュメントでは，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．そのため，その内容を説明するために，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られる．このような場所による単語の関連を表す性質を用いてデータ行列を生成できれば，自動的に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．

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

本章では，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，日本語ドキュメントの全文検索などにおいて広く利用されている形態素解析器 ChaSen[9)]，及び，東京大学中川研究室・横浜国立大学森研究室で開発された専門用語自動抽出システム [10)] を使用する．これらのシステムによって抽出される任意の 2 語が同一のドキュメント中で出現する場合の距離と頻度から各単語間の関連の度合いを求め，章で示す意味の数学モデルに適用することにより，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する．本方式では，検索対象が包含する特定分野，及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，専門用語自動抽出システムの概要を示し，2.2 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．そして，2.3 節では，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の実現方法について示す．

2.1 専門用語自動抽出システム

専門用語自動抽出システム [10)] とは，専門分野のコーパス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．このシステムは，形態素解析器 ChaSen[9)] によるテキストの形態素解析結果を入力として専門用語の抽出を行っている．

まず，形態素解析結果から候補語となる単名詞，及びそれらの複合名詞を抽出する．次に，各候補語の出現頻度と，それらの候補語を形成する単名詞のバイグラムの相乗平均を用いて各候補語の重要度を計算する．ここで，バイグラムとは，2 つの単語が隣接して出現する頻度を表す．重要度計算法の詳細については文献 [11)] に示されている．最後に，抽出された専門用語を重要度が高い順に出力する．この出力には各専門

用語の重要度と出現回数も付加されているため，情報検索において非常に有用である．

2.2 「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．

2.3 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

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，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)]，及び東京大学中川研究室・横浜国立大学森研究室で開発された用語抽出システム [10)] 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抽出することができる．

(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，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．

$$P_{ij}(d) = \frac{F(j,i,d)}{f(i)} \tag{2}$$

ここで $F(i,j,d)$ は w_i が w_j から距離 d の位置に出現する頻度を，$f(i)$ は w_i がドキュメント内で出現した回数を表している．

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．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．

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

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．

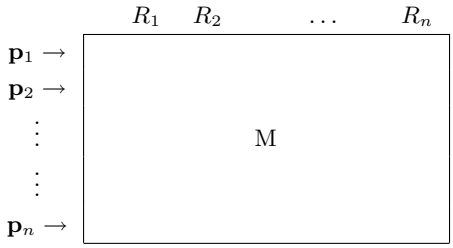

図 1 データ行列 M によるメタデータの表現
Fig. 1 Metadata represented in data matrix M

(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．

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

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．

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

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

本方式の有効性を検証するため，本方式に基づいたメタデータ空間を生成して検索実験を行った．

本方式は，専門家の人手による辞書や用語辞典などが存在しない分野における単語間関連連想検索の実現のためのメタデータ空間生成を想定している．

4.1 実験環境

本実験において，対象となるドキュメントからの単語の抽出には，2.1 節で示した専門用語抽出システムを用いた．このシステムは入力に形態素解析の結果を要求するため，形態素解析器 ChaSen を前処理として併用した．これらの処理の結果得られる語群に対して本方式を適用することにより，メタデータ空間を生成して検索実験を行った．また，ここでは，対象となるドキュメントには Web サイト「 IT 用語辞典 e-Words 」[12] 内の「コンピュータ」「人工知能」「エキスパートシステム」「プログラミング言語」「コンパイラ」「インタプリタ」の 6 ページから説明本文のみを抜き出したものを使用した．この分野には辞書や用語辞典などが無数に存在するが，up-to-date な辞書生成の必要性が認められる分野でもあるため，この分野を対象とすることは本方式の性能評価，及び有効性の検証に適すると考えられる．

4.2 実験結果

検索結果として,「人工知能」,「エキスパートシステム」,「プログラミング言語」の 3 つのコンテクストで検索を行った場合の検索結果をそれぞれ表 1，2，3 に示す．これらは検索結果の上位 10 件を示している．

コンテクスト「人工知能」の場合，検索結果の上位には「画像理解システム」,「意味」,「音声理解システム」,「音声」,「画像」など,「人工知能」と関連の強い語が検索されている．元のドキュメントを参照してみても,「人工知能」の説明にこれらが用いられており，単語間の関連を適切に表現できていると考えられる．

コンテクスト「エキスパートシステム」の場合，検索結果の上位には「エキスパートシステム」,「専門家」,「翻訳」など,「エキスパートシステム」及びこれと関連の強い「人工知能」という語から連想される語が検索されている．それ以下にも「機械翻訳システム」,「専門知識」,「知識データ」などといった語が検索されており，説明文における単語間の関連を適切に表現できているといえる．

コンテクスト「プログラミング言語」の場合，検索結果の上位には「画像理解システム」,「音声理解システム」といった，直感的にもあまり関連のなさそうな語が多く検索されている．そこで，対象となったドキュメントを参照してみると,「プログラミング言語」の説明部分以外では,「人工知能」という語の説明において「プログラミング言語」という語が使用されていることが分かる．使用した全ドキュメント中で「プログラミング言語」という語が出現するのはこの 2 箇所だけであり，これが検索結果に大きく影響を及ぼしたと考えられる．

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

語	相関量
画像理解システム	0.834322
意味	0.815629
音声理解システム	0.804567
音声	0.747101
画像	0.713491
人工知能	0.636652
機械翻訳システム	0.598716
プログラミング言語	0.548544
翻訳	0.484473
コンパイラ	0.444111

4.3 考　　察

コンテクスト「人工知能」,「エキスパートシステム」では，単語の局所性に基づいて各単語間の距離重み頻度関数を求めることにより，対象となるドキュメント中の記述を反映して，単語間の関連を適切に表現できていたと考えられる．

他方，コンテクスト「プログラミング言語」の検索結果から，実際の単語間の関連を必ずしも反映しきれない場合があることも分かる．しかしながら，これはその語とあまり関係のないドキュメントにおける単語間の距離と頻度を忠実に反映した結果であり，その語と関連の強い語に関するドキュメントを蓄積することにより改善が可能であると考えられる．

本実験は，特定分野に関するドキュメント内における単語の局所性を用いることにより，各専門用語間の関連性を反映した検索を実現することが可能であることを示している．

5. おわりに

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．

また，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

表 2 実験結果 1-2 （コンテクスト：エキスパートシステム）
Table 2 Experimental results 1-2 (Context:エキスパートシステム)

語	相関量
エキスパートシステム	0.725869
専門家	0.618273
翻訳	0.438940
問題解決技法	0.433963
機械翻訳システム	0.421494
画像	0.395523
専門知識	0.387459
推論能力	0.385572
知識データ	0.385561
汎用コンピュータ	0.371018

表 3 実験結果 1-3 （コンテクスト：プログラミング言語）
Table 3 Experimental results 1-3 (Context:プログラミング言語)

語	相関量
画像理解システム	0.826678
意味	0.809834
音声理解システム	0.792979
音声	0.736378
画像	0.700741
プログラミング言語	0.621219
機械翻訳システム	0.579008
人工知能	0.532856
実行	0.529440
ソフトウェア	0.524714

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も，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，実際の特定分野全体の内容を包含したドキュメント群を対象とした単語間関連連想検索の実現，及び，その定性的な検索精度の検証が挙げられる．

参 考 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/hiki/ChaSen/
10) http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/resource/termext/atr.html
11) 中川裕志，森辰則，湯本紘彰：“出現頻度と連接頻度に基づく専門用語抽出，” 情報処理学会第 145 回自然言語処理研究会，Vol.10 No.1, pp. 27 - 45, (2003).
12) http://e-words.jp