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

An Integration and Extension Method of a Metadata Space from Indexes of Documents

中西 崇文♥ 岸本 貞弥♦
櫻井 鉄也♠ 北川 高嗣♠

Takafumi NAKANISHI Sadaya KISHIMOTO
Tetsuta SAKURAI Takashi KITAGAWA

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．

In this paper, we present an integration and extension method of metadata spaces from indexes of two or more documents. This method make it possible to integrate metadata spaces based on the relation between words by using common terms in indexes of documents. The number of the vocabularies, which the metadata space constructed from the index of a document can express, is restricted. The problem is solved by this method that uses the document of two or more parts. It is thought that the metadata 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 In this paper, we also present an implementation method for applying our method to words related associative search. We clarify effectiveness of our method by several experiments.

♥ 学生会員 筑波大学大学院システム情報工学研究科
takafumi@nalab.is.tsukuba.ac.jp
♦ 非会員 筑波大学大学院理工学研究科
kishimoto@nalab.is.tsukuba.ac.jp
♠ 非会員 筑波大学大学院システム情報工学研究科
{sakurai,takashi}@is.tsukuba.ac.jp

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，文献[1][2][3]で言葉と言葉の関係の計量による検索機構である意味の数学モデルを提案している．意味の数学モデルでは，検索対象をベクトル化し，言葉と言葉の関係を計量する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コンテクストに応じてメタデータ空間における部分空間を動的に選択し，射影して計量することにより検索対象の検索を可能とする．

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．メタデータ空間は基本データとよばれる特徴付きベクトルを要素とした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．

これまでのメタデータ空間の生成方式として，Longman Dictionary of Contemporary English(以下，Longman)[4]という英英辞典を用いる方法[2], 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メタデータ空間を生成する方式[5],[6]が提案されている．しかしながら，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[2][5][6]では，見出し語を特徴づけする語(特徴語)を選定する必要性があるなど，自動化が難しく，高い専門性を要する．また，これらの方式[2][5][6]は，対象とする分野，もしくはそれに近い分野の用語辞典を有することを前提にしており，これらの用語辞典が存在しない分野については言及していない．学術分野においては，その分野の用語辞典が存在する可能性は高いが，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．

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[7]について研究を進めてきた．この方式は，対象とする特定分野について書かれた書籍の索引に注目する．索引に挙げられている語は，著者や編者が書籍の内容，つまり書籍に記述されている対象とする特定分野の中のキーワードとして抽出した語であると考えられる．よって，それらの語はその特定分野における，基本的かつ重要な語であると考えられる．また，一般的に書籍には読者が理解しやすいように，関係のある内容が近くにまとめて書かれていることから，その内容に関連する語がかたまって現れやすく，書籍の中で近くに書かれている語同士は関連性が高いと考えられる．そして，それぞれのページ番号は，語が出現する場所情報を示すIDであるとみなすことができる．これらのことから，索引に挙げられている語を特徴語とみなし，ページ番号を特徴語によって特徴づけることによって行列を生成することにより，書籍の構成を加味することによって語同士は関連性を加味したメタデータ空間を生成できる．つまり，書籍によってメタデータ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，メタデータ空間の生成ができる．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あり，用語辞典に収録されている語数よ

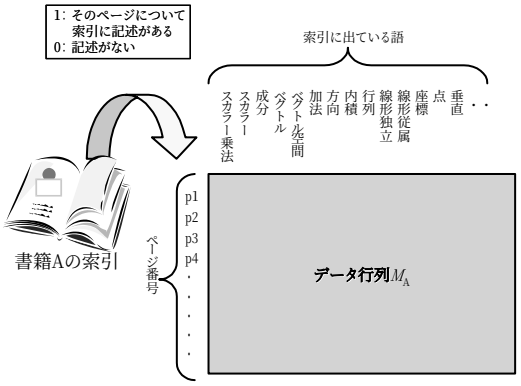

図 1 書籍の索引部によるデータ行列.
Fig.1 Data matrix M by index of a book.

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．

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．本稿は複数の書籍の索引を用いることにより，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．

2. 意味の数学モデルの基本構成

本節では，様々な単語(以下，印象語)によって表現した問い合わせに対応したメディアデータを検索することを目的とした意味の数学モデルの概要を示す．詳細は，文献[1][2][3]に述べられている．

(1) メタデータ空間 *MDS* の設定
検索対象となるメディアデータをベクトルで表現したデータをマッピングするための正規直交空間(以下，メタデータ空間 *MDS*)を設定する．本稿では，このメタデータ空間 *MDS* を複数の書籍の索引を用いることによって生成する方式について提案している．

(2) メタデータをメタデータ空間 *MDS* へ写像
設定されたメタデータ空間 *MDS* へメディア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，同じ空間に検索対象データのメタデータが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意味的な関係を空間上でのノルムとして計算することが可能となる．

(3) メタデータ空間 *MDS* の部分空間(意味空間)の選択
検索者は与える文脈を複数の単語を用いて表現する．検索者が与える単語の集合をコンテクストと呼ぶ．このコンテクストを用いてメタデータ空間 *MDS* に各コンテクスト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 *MDS* 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(以下，重み)を持つ軸からなる部分空間(以下，意味空間)が選択される．

(4) メタデータ空間 *MDS* の意味空間における相関の定量化
選択されたメタデータ空間 *MDS* の部分空間(意味空間)において，メディア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，与えられたコンテクストと各メディアデータとの相関の強さを定量化している．この意味空間における検索結果は，各メディア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．

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，複数の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を示す．本方式は，それぞれの書籍の索引でデータ行列を作成し，そのデータ行列群を統合することによって空間の拡張統合を実現する．3.1 節では，書籍の索引からデータ行列を生成する方式を示す．3.2 節では 3.1 節の方式で生成されたデータ行列群を統合する方式を示す．

3.1 書籍の索引部を用いたデータ行列生成

メタデータ空間を生成するためには，データ行列を生成する必要がある．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を示す．本方式の詳細は文献[7]で報告している．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．

(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付ける．

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

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 $n + R$ 列のデータ行列 M_1 を生成できる．ここで，R は章，節番号を特徴として付け加えた分である．

3.2 複数のデータ行列の統合

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 3.1 節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，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 A の索引から生成されたデータ行列 M_A, 書籍 B の索引から生成されたデータ行列 M_B, を対象として統合したデータ行列 M 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．
なお，文献[6] 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[6] のベクトル要素の統合もない．

（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 A」「書籍 B」を特徴語として，書籍 A の索引のページ番号の場合「書籍 A」に ”1”，書籍 B の索引のページ番号の場合「書籍 B」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. 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．本実験では，「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】」(以下，シスアドの教科書)[8]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」(以下，DB の教科書)[9]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．これらの空間を IT 分野に関する画像メディアデータを対象として，画像メディアデータ検索に適用した．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，IT 用語辞典)[10] 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．

実験 A では, シスアドの教科書のみで生成したメタデータ空間での画像メディア検索と，シスアドの教科書の索引部と DB の教科書の索引部を統合したメタデータ空間での画像メディア検索を比較した．

実験 B では，従来の時点によって生成する方式として IT 用語辞典からメタデータ空間[4][5]を生成した画像メディア検索と，本方式である複数の書籍の索引によって生成されたメタデータ空間での画像メディアデータ検索が，どの程度結果が一致するかを比較行った．これにより，
本方式が適切な用語辞典が存在しない分野における空間生成の代替方式として利用可能かを検証した．

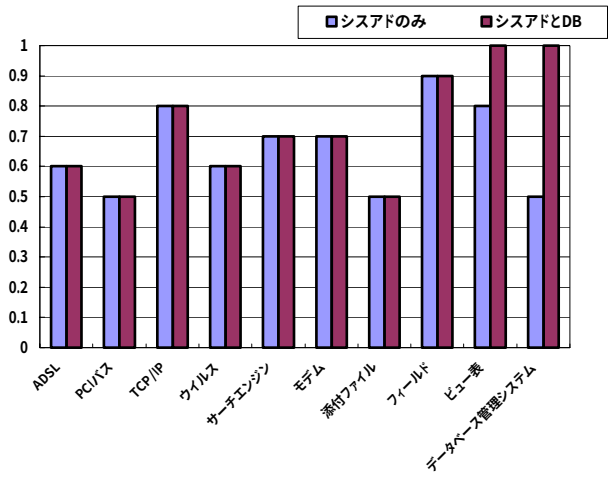

図 2 実験 A の結果.
Fig.2 Experimental results A.

4.1 実験 A

IT 分野に関する画像メディアデータを対象として, 画像メディアデータ検索に適用し，シスアドの教科書の索引部のみで生成したメタデータ空間と，シスアドの教科書と DB の教科書の索引部で生成したデータ行列を統合し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．

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

なお，システムの検索結果出力数は上位 10 位とする．

結果を図 2 に示す．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．

4.2 実験 B

IT 分野に関する画像メディアデータを対象として, 複数の書籍の索引部によるメタデータ空間の検索の場合が用語辞典によるメタデータ空間の検索の場合と比べてどれくらい一致しているかを評価することにより，本方式が適切な用語辞典が存在しない分野における空間生成の代替方式として利用可能かを検証した．

複数の書籍の索引部によるメタデータ空間として，実験 A で構築したシスアドの教科書と DB の教科書の索引部を使用した．また，辞典として IT 用語辞典を用いた．なお，検索対象となる画像メディアデータは実験 A と同様 418 個を使用した．

結果を図 3 に示す．平均して上位 5 位では 7 割，上位 10 位では 6 割程度検索結果が一致している．これらの一致は，パターンマッチングとしての一致だけではなく，画像メディアデータのメタデータに検索語が含まれてなくても出力し，一致した例がほとんどであった．このことから，容易に少ない手間で用語辞典による方式と似た，語と語の関係に合致し

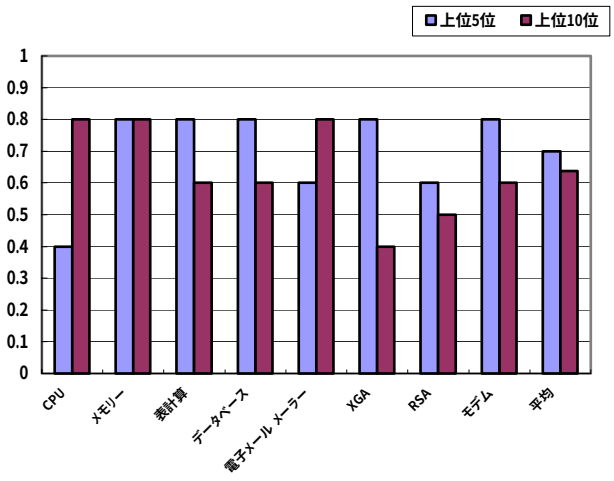

図 2 実験 B の結果.
Fig.2 Experimental results B.

た検索結果を得られると考えられる.

5. まとめと今後の課題

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, 意味的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, 意味的連想検索の導入が容易に可能になると考えられる.

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] 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).

[7]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).

[8]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).

[9]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).

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

中西 崇文 Takafumi NAKANISHI
筑波大学大学院システム情報工学研究科在学中. 2001 年筑波大学第三学群情報学類を卒業. マルチメディアシステムに関する研究に興味を持つ. 情報処理学会学生会員. 電子情報通信学会学生会員. 日本データベース学会学生会員.

岸本 貞弥 Sadaya KISHIMOTO
筑波大学大学院理工学研究科在学中. 数理ソフトウエア利用支援の研究に興味を持つ.

櫻井 鉄也 Tetsuya SAKURAI
1986 年名古屋大学院工学研究科博士課程前期課程情報工学専攻修了. 同年同大学助手. 筑波大学講師を経て, 現在, 筑波大学電子・情報工学系助教授. 工学博士. 非線形方程式の解法と有利関数近似値法の応用, および数理ソフトウエアの利用支援の研究に従事. 1996 年日本応用数理学会論文賞受賞. 日本応用数理学会会員.

北川 高嗣 Takashi KITAGAWA
筑波大学電子・情報工学系教授. 1978 年名古屋大学工学部卒業. 1983 年同大学院工学研究科博士過程修了. 工学博士. スタンフォード大学計算機科学科客員研究員, 愛媛大学理学部数学科講師, 筑波大学電子・情報工学系助教授を経て現在に至る. 数値解析, 逆問題, マルチメディア情報システムの研究に従事. 日本応用数理学会会員.