

事象データ間の因果関連性計量機能をともなったベクトル空間検索方式

図子泰三[†] 清木 康^{††} 鷹野孝典[†]
波内みさ^{†††} 但田育直^{†††}

本稿では、検索者の与える検索語を1つの事象としてとらえることが可能な領域を対象として、事象間の因果関連性をともなったベクトル空間検索方式を示す。事象データをベクトルとして表現することによって、事象データ間の因果関係を数値として記述することができる。因果関係を数値で表現するメリットとして、特定の事象データに自由に重要度を設定することができ、また因果関連性計量をともなった文書検索にも応用できる。擬似的な事象データ・文書データによる実験により、提案方式の有効性を確認し、宇宙工学分野の実文書データを対象とした実験により、提案方式の実現可能性を明らかにする。

A Vector Space Retrieval Method with Causal Relationship Computation Functions for Event Data

TAIZO ZUSHI,[†] YASUSHI KIYOKI,^{††} KOSUKE TAKANO,[†]
MISA NAMIUCHI^{†††} and IKUNAO TADA^{†††}

In this paper we propose a vector space retrieval method with causal relationship computation functions for event data. The method is used for the domain in which a query given by a user can be regarded as an event. By expressing an event as a vector, causal relationships between events can be described as numerical values. This feature enables to set a level of importance on a specific event flexibly. The method can be applied to document retrieval with causal relationship computation. We clarify feasibility and effectiveness of the method by showing several experimental result using synthetic and real event data and document data.

1. はじめに

近年、様々な組織内において、大量の文書データが生成され、それらがデータベースに格納されている。また、それらの文書データを活用するために、組織ごとに検索エンジンが構築されている。データベースや情報検索の研究分野では、これらの文書データ群を対象とした検索方式として、ベクトル空間モデルによる検索方式が有効であることが確認されている。

従来のベクトル空間モデルでは、対象とするドキュメントデータ群の中に出現する各単語をベクトルデータとして表現し、単語間の意味的な同義性や類似性を

計算可能な計量系を提供している。しかし、ある検索対象領域において、検索者が検索語として与える単語を事象(event)としてとらえることが可能な場合、事象間の類似性を計量するといった検索要求を満たした計量系だけでは十分ではなく、事象間の因果関係を計量可能な系が必要となる。

情報検索の分野で提案されているベクトル空間検索方式である SMART システム¹¹⁾ や Latent Semantic Indexing (LSI)⁴⁾ では、文書群を対象とした検索空間を生成するために、索引語・文書行列を作成する。索引語・文書行列では、1つの文書データは、その文書に出現する単語を要素に持つベクトルとして表現することによって、検索者の与える検索語と同義的・類義的関連度の高い文書データを検索することが可能となる。しかし、これらの方式では、検索対象領域における事象間の因果関係を計量可能な索引語・文書行列を作成することは困難であり、検索者の検索語と因果関係のある文書データを検索することはできない。

[†] 慶應義塾大学政策・メディア研究科
Graduate School of Media and Governance, Keio University

^{††} 慶應義塾大学環境情報学部
Faculty of Environmental Information, Keio University

^{†††} 宇宙航空研究開発機構
Japan Aerospace Exploration Agency

本稿では、検索対象領域における事象間の因果関係を計量可能なベクトル空間検索方式を示す。提案方式では、事象ベクトル空間生成方式と事象ベクトル検索方式の2方式から構成されている。事象ベクトル空間生成方式では、従来のベクトル空間モデルと異なり、検索対象領域において発生しうる事象データ間の因果関係を、その分野の専門家の判断によってベクトルデータとして定義する。これによって、事象間の因果関連性が計量可能な検索空間の構築が可能となる。事象ベクトル検索方式では、定義された事象ベクトル間の相関度を計量するために、従来の情報検索の分野で適用されている内積計算やコサイン尺度²⁾を使用して実現する。さらに、事象ベクトル間の相関度計量に意味の数学モデル^{8),9)}による意味的連想処理機構を適用することによって、検索者の文脈に応じた因果間関連性計量の実現が可能となる。

因果関連性を計量する方式は、これまでに様々な分野で研究されてきた。一般に広く知られているものとしては、事象間の因果関係に確率モデルを適用したベイジアンネットワーク¹⁰⁾、同様に統計的処理を適用した構造方程式モデリング¹⁾、共分散構造分析³⁾などがある。これらの従来手法と比較して、提案方式の特徴、つまり、事象間の因果関連性をベクトル空間を用いて計量することの特徴として次の3点をあげることができる。

- (1) ベクトルデータに対する様々な計算手法を適用可能
事象データ群をベクトルデータ化しておくことにより、内積計算、コサイン尺度、意味の数学モデルなどの計量手法を適用することができ、ベクトル空間上での事象間の因果関係を対象とした検索が実現可能となる。さらに、クラスタリング手法^{7),14)}を適用することによって、事象データ群の因果関連性に応じた分析が可能となる。
- (2) 特定の事象データを重要視した検索の実現
検索者の与える検索語事象ベクトルの特定の要素に、自由に重みを設定することによって、特定の事象の重要度を変化させることができる。これによって、1つの検索語事象に対して、複数個のベクトル表現を設定することができ、検索者の目的やアプリケーションに応じた因果関連性検索が可能となる。このことによって、因果関連性計量をともなった検索エンジンにおける personalization 機能の実現が可能であると考えられる。

- (3) 因果関連性計量をともなった文書検索を実現
文書データに事象データをメタデータとして付与すれば、それらの事象メタデータベクトルの合成したものを文書ベクトルとして表現することができる。さらに、このように表現される文書ベクトル群と、検索者が与える検索語事象ベクトルとの相関度計量を行うことによって、因果関連性計量をともなった文書検索が可能となる。たとえば、ある事象Aの原因となりうる事象について記述された文書の検索を実現できる。

本稿では、提案方式である因果関連性計量をともなったベクトル空間検索方式を示し、擬似的な事象データ群・文書データ群を用いた実験結果を示すことによって、提案方式の有効性を確認する。さらに、宇宙工学分野の実文書データを対象とした実験結果により、提案方式の実現可能性を明らかにする。

2. 因果関連性を扱うベクトル空間生成方式

本章では、検索対象領域における事象間の因果関連性が計量可能なベクトル空間を生成する方式を示す。ベクトル空間の生成は、ベクトル空間生成用行列を作成するフェーズとそのベクトル空間生成用行列を用いて実際に空間を生成するフェーズから成る。これらの2フェーズは、次のように構成される。

2.1 ベクトル空間生成用行列の作成

事象データ間の因果関係を計量するためのベクトル空間行列として、3つの特徴行列 M_c , M_e , M を生成する。これらの行列の生成ステップを次に示す。この説明において、*feature* (特徴語)とは、ベクトル空間が扱う領域を表現(定義)するのに必要な単語の集合であり、基本データとは、ベクトル空間が扱う領域中の単語の集合である。

Step-I *feature* (特徴語)群, 基本データ群の設定
3つの行列 M , M_c , M_e に対し, n 個の事象群 $e_1 \sim e_n$ を *feature*, 基本データとして, それぞれ行列の横軸, 縦軸に設定し, $n \times n$ の正方行列を生成する。

Step-II 特徴付けの設定

Step-I で *feature* 群, および基本データ群を設定した3つの行列 M , M_c , M_e について, それぞれ特徴付けの設定を行う。それぞれの行列に対する特徴付けの設定方法を以下に示す。

M : 基本データ中の事象単語について, *feature* の中から, 自分自身を表す事象単語, 直接的な原因である事象単語, および直接的な結果である事象単語に“1”を設定(相関があるこ

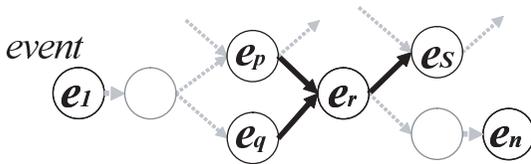


図 1 事象データ群
Fig.1 An event data set.

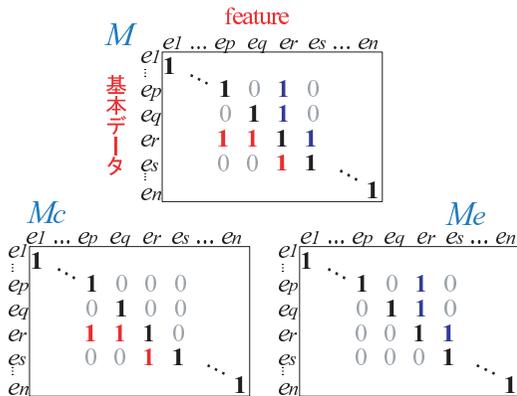


図 2 ベクトル空間生成用行列 M , M_e , M_c
Fig.2 Vector space matrices M , M_e , M_c .

とを“1”として表現)し、それ以外の事象単語に“0”を設定(相関がないことを“0”として表現)する。図1の事象 e_r を例とすると、自分自身である e_r と、直接的な原因となる事象である e_p, e_q 、直接的な結果となる e_s の4つの事象に1を付与する。この操作を基本データ中の全事象単語に適用し、 M (図2上部)を生成する。

M_c : 基本データ中の事象単語について、*feature*の中から、自分自身を表す事象単語、および直接的な原因である事象単語に“1”を設定し、それ以外の事象単語に“0”を設定する。同様に、図1の事象 e_r を例とすると、自分自身である e_r と、直接的な原因となる事象である e_p, e_q の計3つの事象に“1”を付与する。この操作を基本データ中の全事象単語に適用し、 M_c (図2下部左)を生成する。

M_e : 基本データ中の事象単語について、*feature*の中から、自分自身を表す事象単語、および

実際には、図1のように事象間の因果関係がグラフによって示されている領域は多くないと考えられる。現実的な手法としては、直接的な原因(または結果)であるという判断は、対象としている領域の専門家の判断により行うことになる。もしくは、その領域内で実際に起こった事象のログ、つまり、事例をベースにする。

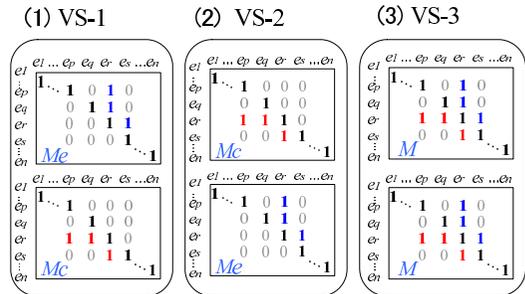


図 3 ベクトル空間生成用行列の組(上段: 検索対象データベクトル形成のための行列, 下段: 問合せ語ベクトル形成のための行列)
Fig.3 Combinations of the vector space matrices.

直接的な結果である事象単語に“1”を設定し、それ以外の事象単語に“0”を設定する。同様に、図1の事象 e_r を例とすると、自分自身である e_r と、直接的な結果となる事象である e_s の計2つの事象に“1”を付与する。この操作を基本データ中の全事象単語に適用し、 M_e (図2下部右)を生成する。

2.2 ベクトル空間生成用行列の組の設定

2.1 節で生成した3つのベクトル空間行列 M, M_c, M_e を用いて、「ある事象の原因である事象群」について検索する、「ある事象の結果である事象群」について検索するなどの、検索目的に応じたベクトル空間行列の組 $VS-1 \sim VS-3$ (図3)の設定方法について述べる。

- (a) ある事象の原因である事象群を検索する目的の場合
検索対象データベクトル形成のためのベクトル空間行列として M_e を、問合せ語ベクトル形成のためのベクトル空間行列として M_c を設定する。このベクトル空間行列の組を $VS-1$ とする。
- (b) ある事象の結果である事象群を検索する目的の場合
検索対象データベクトル形成のためのベクトル空間行列として M_c を、問合せ語ベクトル形成のためのベクトル空間行列として M_e を設定する。このベクトル空間行列の組を $VS-2$ とする。
- (c) ある事象の原因である事象群、および結果である事象群の両方について検索する目的の場合
検索対象データベクトル形成のためのベクトル空間行列、問合せ語ベクトル形成のためのベクトル空間行列ともに M を設定する。このベクトル空間行列の組を $VS-3$ とする。

以上設定したベクトル空間行列の組($VS-1 \sim VS-3$)

を、通常のベクトル間の内積計算・コサイン尺度による検索方式、および、意味の数学モデルによる意味的連想検索方式といったベクトル空間モデルによる検索方式に適用することによって、「ある事象の原因である事象群」について検索する、「ある事象の結果である事象群」について検索するなど、事象間の因果関係に関する検索をベクトル空間によって実現することが可能となる。

2.3 ベクトル空間の生成

2.1 節に述べたベクトル空間生成用行列を用いて検索用ベクトル空間を生成する方法について示す。検索用ベクトル空間の生成については、次章で述べる事象ベクトルデータ間の相関度計量にどの手法を適用するかによって異なってくる。

事象ベクトルデータ間の相関度計量を行う際に、単純に内積計算やコサイン尺度²⁾を適用する場合は、2.1 節において作成した行列データをそのまま利用すればよい。すなわち、 n 個の事象データを対象とした領域では、行列 M (もしくは M_c, M_e) によって表現される n 次元のベクトル空間を検索ベクトル空間とする。

意味の数学モデルで提案されている手法^{8),9)}では、2.1 節において作成した行列データ M (もしくは M_c, M_e) に対して相関行列 $M^T M$ (もしくは $M_c^T M_c, M_e^T M_e$) を固有値分解し、算出された固有ベクトルを次元とする正規直交化ベクトル空間を生成する。このように正規直交化ベクトル空間を生成することによって、データ間の距離(相関度)を計量する検索系を構築することが可能となる。

2.4 検索対象文書ベクトルの生成

検索対象となる文書データをベクトル化する方法を示す。文書データは $e_1 \sim e_n$ のうち 1 個以上の事象がメタデータとして付与されていることを前提とする。これらの各文書データに付与されているメタデータ(事象)をそれぞれ和(OR)演算によって合成し、文書ベクトルを構築する。合成するベクトルは、原因事象検索の場合は、 M_e において表現されているベクトルを、結果事象検索の場合は、 M_c において表現されているベクトルをそれぞれ使用する。つまり、検索目的に応じて、文書検索のベクトル表現も異なることになる。

3. 因果関連性計量をともなったベクトル空間検索方式

本章では、2 章で述べた方式によって生成したベクトルデータを対象として、因果関連性計量をともなったベクトル空間検索方式を示す。

2 章で示したとおり、各事象ベクトルは、直接的な原因(あるいは、結果)だけに値が付与されたベクトルとなっている。このことは、ある事象で検索したときに、この事象の直接的に因果関係のある事象しか検索できないことを示している。ここでは、検索者の検索語として与えられる事象に対して、間接的な因果関係まで含めた検索を実現するために、検索語事象ベクトルの拡張方式を示す。さらに、拡張された検索語事象ベクトルと、検索対象事象ベクトルとの相関度計量について示す。

3.1 検索語ベクトルの拡張

2 章で示したように、検索語として使用する事象ベクトルと検索対象として使用する事象ベクトルの両方も直接的な因果関係のある事象にのみ“1”が付与されている。これらのベクトルデータを使用して検索を行っても、直接的な因果関係しか計量することができない。ここでは、間接的な因果関係にまで及ぶ検索を可能とするために、検索者が与える検索語事象ベクトルを拡張する方式を示す。

検索者が与える検索語事象ベクトルを拡張するアルゴリズムを以下に示す。

- (1) 与えられた事象ベクトルの要素のうち“1”が付与されている事象(つまり直接的な因果関係のある事象)を列挙する。
- (2) 列挙したそれぞれの事象について、事象ベクトルの“1”が付与されている要素に拡張関数 f_{expand} を適用する(つまり、事象ベクトルの要素のうち“1”が付与されている要素は $f_{expand}(1)$ に変更する)。
- (3) (2)における各ベクトルを元の検索語事象ベクトルに足し合わせる。

(1)~(3)の操作を再帰的に適用し、再帰のフェーズが拡張定数を超えるまで続ける。拡張関数 f_{expand} としては、アプリケーションに応じて様々な関数を設定することが可能であるが、一般的には次のような関数を適用する。

$$f_{expand}(x, phase) = x * k^{phase}$$

x は対象としているベクトルの要素の値、 k は任意の定数、 $phase$ は再帰のフェーズを表す。より間接的な事象ほど因果関連性を低くするには、 $k < 1$ とし、逆に、より間接的な事象ほど因果関連性を高くするには、 $k > 1$ と設定すればよい(前者が一般的である)。また、 $phase$ は、初期値が“0”で、再帰するごとにインクリメントする。

拡張定数は、検索語ベクトルの拡張をどこまで行う

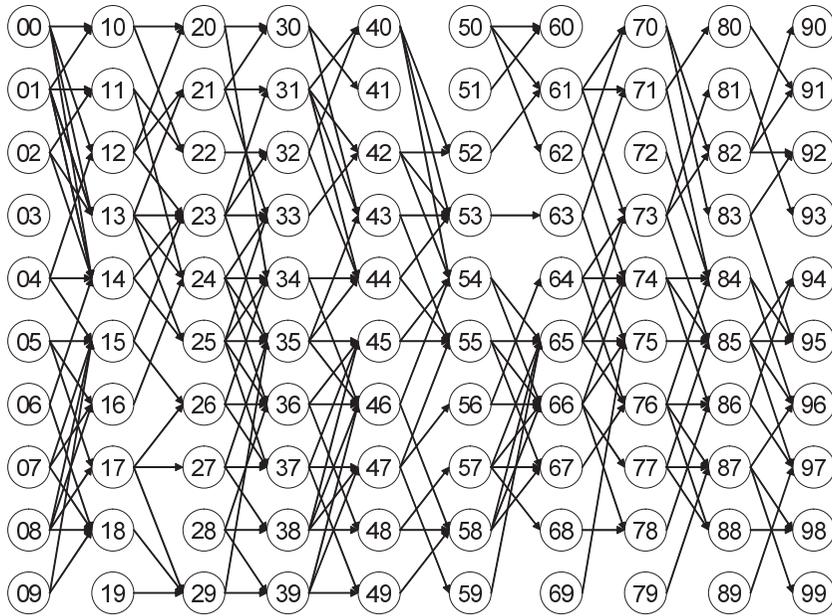


図 4 実験用擬似事象データ

Fig. 4 A synthetic event data set for the experiments.

かを決定する定数である．たとえば，2つ離れた間接的な事象まで因果関連性を計量したい場合には，拡張定数を“2”にすればよい．すべての間接的な因果関連性を有する事象を検索する場合は，拡張定数を設定する必要はない．

3.2 因果関連性に基づいたベクトル間の相関度計量

実際に検索を行う際には，検索者が与える検索語事象ベクトルに対して，3.1節で示した方法によってベクトルを拡張し，この拡張した検索語ベクトルと検索対象事象ベクトルとの相関度を計量する．相関度を計量する方法としては，ベクトル空間モデルにおいて一般的に使用されている内積計算やコサイン尺度を適用することによって実現できる．さらに，検索語事象ベクトル・検索対象事象ベクトルに対して，意味の数学モデル^{8),9)}による意味的連想検索方式を適用することも可能である．

4. 実験

提案方式の有効性および実現可能性を検証するための実験について示す．

4.1 実験環境

実験用データとして，擬似的な事象データ群と文書データ群を生成した．

4.1.1 擬似事象データの生成

本実験で使用するための擬似事象データを100個生成した．これらの擬似事象データは e_{ij} という形式の

IDを持っている． i, j はそれぞれ0から9のいずれかの値である．つまり，これらの100個の事象データは e_{00} から e_{99} ということになる．

作成した100個の事象間の因果関係を次に示すルールに基づいて決定した．

- (1) e_{pq} の結果として起こる事象は， $e_{(p+1)j}$ のうちのいずれかである (e_{pq} の原因として起こる事象は， $e_{(p-1)j}$ のうちのいずれかである)．
- (2) 1個の事象から結果として起こる事象は5個以下である．

これらのルールに基づいて作成した事象データ，および，事象データ間の因果関係をグラフにしたものを図4に示す．ルール(2)に関しては，乱数を発生させて決定した．

4.1.2 擬似文書データの作成

実験で使用する20件の擬似文書データを表1に示す．これらの文書データは，実験結果の解釈を容易にするために，次のように意図的に偏ったメタデータを付与している．

doc01～05 事象52の原因について記述された文書

doc06～10 事象52の結果について記述された文書

doc11～15 事象56の原因について記述された文書

doc16～20 事象56の結果について記述された文書

たとえば，doc01のメタデータ(40 42 31 33 11)はすべて，図4中の事象52から原因の方向にたどっていくと現れる事象となっている．

4.1.3 実験システム

4.1.1 項で作成した擬似事象データより、2章で示した方法によって、行列 M_c, M_e を生成した。さらに、原因事象検索のための行列の組 $VS-1$ 、結果事象検索のための行列の組 $VS-2$ をそれぞれ設定した。また、4.1.2 項で作成した擬似文書データに関しては、2.4 節

で示した方法に基づいて文書ベクトルを生成した。生成した事象ベクトル・文書ベクトルをテキストファイルの形式でデータベースを構築した。

3章で示した検索語ベクトル拡張方式、および、ベクトル間の相関度計量系を、C 言語を用いて実装した。検索語ベクトル拡張方式では、次の拡張関数を採用した。

$$f_{expand}(x, phase) = x * 0.5^{phase}$$

また、拡張定数は設定していない。ベクトル間の相関度計量には内積計算を使用した。

実験システムを構築した計算機の CPU は Intel Pentium-M 1300 MHz、メモリは 750 MB、OS は Windows XP である。

4.2 実験 A

[方法・結果]

実験 A では、4.1.1 項で作成した 100 個の擬似データ群に対して提案方式を適用し、適切な因果関連性計量が実現できているかを確認する。具体的には、事象 e_{44} を検索語とし、原因事象検索・結果事象検索の 2 種類の検索を行った。図 4 のうち、事象 e_{44} と因果関係のある事象だけを抽出したものを図 5 に示す。検索結果を表 2 に示す。

[考 察]

表 2 と図 5 より、事象 44 を対象とする原因事象検

表 1 擬似文書データ

Table 1 Synthetic document data.

文書 ID	メタデータ					
doc01	40	42	31	33	11	事象 52 の原因について記述された文書
doc02	40	32	22	11	02	
doc03	42	33	22	25	15	
doc04	42	32	10	15	07	
doc05	31	21	24	10	16	
doc06	61	70	73	85	91	事象 52 の結果について記述された文書
doc07	61	72	80	84	97	
doc08	61	81	82	85	96	
doc09	70	71	83	84	93	
doc10	73	83	85	92	97	
doc11	47	38	26	28	19	事象 56 の原因について記述された文書
doc12	47	37	29	13	01	
doc13	47	39	15	17	19	
doc14	37	39	26	17	18	
doc15	39	13	15	05	08	
doc16	64	66	77	82	85	事象 56 の結果について記述された文書
doc17	64	74	78	81	85	
doc18	66	78	82	84	90	
doc19	73	87	88	96	99	
doc20	66	84	85	88	97	

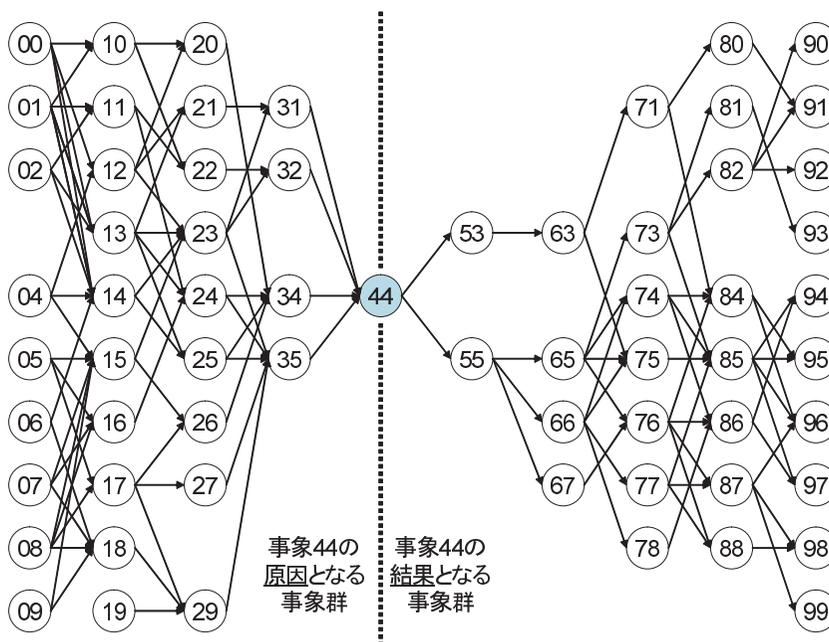


図 5 事象 44 の因果関係

Fig. 5 Causal relationship of event 44.

索の結果では、図 5 における「事象 44 の原因となる事象群」が検索されており、結果事象検索では、図 5 における「事象 44 の結果となる事象群」が検索されていることが確認できる。さらに、事象 44 の直接的な原因（または結果）ほど高い相関度をともなって上位に検索される傾向があることが分かる。原因事象検索において、間接的な原因である事象 23 が上位に検索されている理由としては、多くの直接的な事象（事象 31, 32, 35）の原因となっているためである。このように、本方式は、より多くの事象の原因となりうる事象をより重要な事象と判断し、高い相関度をともなって上位に検索されるようなメカニズムになっている。他の事象で検索を行った場合にも、この結果と

同様の傾向を持つ結果が得られる。実験 A より、提案方式を用いることによって、事象データを対象として、因果関連性計量による相関度をともなった検索が可能となることを示している。

4.3 実験 B

[方法・結果]

実験 B では、提案方式が文書検索に適用可能であることを確認する。これは、1 章で述べた提案方式の特徴 (3) の検証である。表 1 に示す 20 件の擬似文書データを対象として、次の 4 つの検索を行った（カッコ内は適合（正解）文書の ID）。

- (1) 事象 52 の原因事象検索（適合文書：doc01～05）
- (2) 事象 52 の結果事象検索（適合文書：doc06～10）
- (3) 事象 56 の原因事象検索（適合文書：doc11～15）
- (4) 事象 56 の結果事象検索（適合文書：doc16～20）

検索 (1), (2) の実験結果を表 3 に、検索 (3), (4) の実験結果を表 4 に示す。

[考察]

表 3, 4 より、すべての検索結果において、上位 5 位以上にそれぞれの検索語に対応する適合文書が検索されていることを確認できる。さらに、検索語事象データと直接的な因果関係を持つ事象がメタデータとして付与されている文書ほど上位に検索される傾向がある

表 2 実験 A の結果

Table 2 Result of experiment A.

順位	原因事象検索		結果事象検索	
	ID	相関度	ID	相関度
1	23	0.552859	63	0.494451
2	34	0.407850	55	0.416958
3	32	0.407850	53	0.416958
4	24	0.394899	75	0.364838
5	22	0.374634	67	0.349630
6	25	0.360492	76	0.291358
7	13	0.355409	66	0.285472
8	35	0.315919	65	0.285472
9	31	0.315919	74	0.260599
10	17	0.309058	77	0.247226
11	12	0.309058	85	0.234539
12	29	0.305888	44	0.208479
13	20	0.305888	86	0.203951
14	27	0.264906	78	0.201859
15	26	0.264906	71	0.201859
16	21	0.264906	84	0.182419
17	15	0.254906	73	0.174815
18	14	0.254906	88	0.168216
19	11	0.254906	87	0.145679
20	10	0.254906	81	0.123613
21	44	0.203925	80	0.123613
22	19	0.187317	96	0.101975
23	18	0.187317	82	0.100929
24	16	0.187317	98	0.084108
25	08	0.177705	94	0.084108
26	00	0.177705	91	0.084108
27	07	0.154529	97	0.072840
28	05	0.154529	95	0.072840
29	04	0.154529	99	0.061806
30	02	0.154529	93	0.061806
31	01	0.154529	90	0.061806
32	09	0.127453	92	0.050465
33	06	0.127453	89	0
34	99	0	83	0
35	98	0	79	0
..	(以下 100 位まですべて相関度 0)			

本方式の中で使用する拡張関数のパラメータをチューニングすることによって重要度をコントロールすることは可能である。

表 3 事象 52 の文書検索結果

Table 3 Document search result of event 52.

順位	原因事象検索		結果事象検索	
	文書 ID	相関度	文書 ID	相関度
1	doc04	4.4375	doc08	4.125
2	doc01	4.375	doc06	3.875
3	doc05	4	doc07	3.375
4	doc03	3.875	doc10	3
5	doc02	3.4375	doc09	2.875
6	doc15	1.5	doc19	2.125
7	doc12	1.4375	doc18	2.125
8	doc13	0.375	doc20	1.625
9	doc20	0	doc16	1.5
10	doc19	0	doc17	1
11	doc18	0	doc15	0
12	doc17	0	doc14	0
13	doc16	0	doc13	0
14	doc14	0	doc12	0
15	doc11	0	doc11	0
16	doc10	0	doc05	0
17	doc09	0	doc04	0
18	doc08	0	doc03	0
19	doc07	0	doc02	0
20	doc06	0	doc01	0

表 4 事象 56 の文書検索結果
Table 4 Document search result of event 56.

順位	原因事象検索		結果事象検索	
	文書 ID	相関度	文書 ID	相関度
1	doc11	4.375	doc17	6
2	doc12	3.8125	doc20	5.625
3	doc13	3.625	doc16	5.5
4	doc14	3	doc18	4.625
5	doc15	2.625	doc10	4
6	doc03	1.125	doc19	3.75
7	doc04	0.5625	doc06	3.625
8	doc02	0.3125	doc08	3.375
9	doc20	0	doc07	1.875
10	doc19	0	doc09	1.625
11	doc18	0	doc15	0
12	doc17	0	doc14	0
13	doc16	0	doc13	0
14	doc10	0	doc12	0
15	doc09	0	doc11	0
16	doc08	0	doc05	0
17	doc07	0	doc04	0
18	doc06	0	doc03	0
19	doc05	0	doc02	0
20	doc01	0	doc01	0

ことも確認できる．たとえば，doc01 と doc05 を比較した場合，doc01 には事象 52 の直接的な原因となる事象（事象 40・42）がメタデータとして付与されているが，doc05 には間接的な原因となる事象しかメタデータとして付与されていないため，doc01 のほうが doc05 より高い相関度をともなって上位に検索されている．

また，各検索結果の 6 位～8 位には，弱い（0 でない）相関度を持つ文書が検索されていることを確認できる．これは，6 位以下に検索された文書に付与されているメタデータのうち，ほとんどのメタデータは検索語事象と因果関係はないが，一部のメタデータに関しては，検索語事象と間接的な因果関係が存在することを示している．たとえば，事象 52 の原因事象検索（検索（1））を行った結果の 6 位にランクされている doc15 に関しては，メタデータとして事象 5 が付与されている．図 4 を参照すると，事象 5 は事象 52 の間接的な原因となっているので，5 位以上の文書と比較して弱い相関度をともなって 6 位にランクされていると解釈できる．このことは，検索語事象の直接的でない原因（または結果）に関して記述されている文書は，検索結果として中位にランクされ，本方式によって因果関係の強度を考慮した検索が可能となっている．

4.4 実験 C

[方法・結果]

実験 C では，事象データに対して重要度を導入し，特定の事象の重要度を変化させた場合に，それに応じ

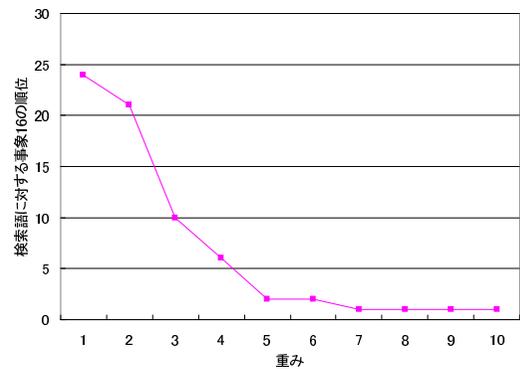


図 6 重みに対する順位の変化

Fig. 6 Ranking transition for the weighting.

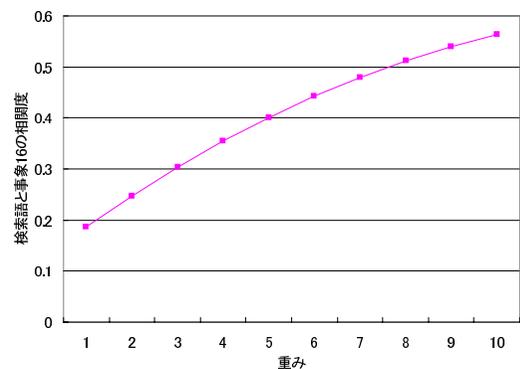


図 7 重みに対する相関度の変化

Fig. 7 Correlation transition for the weighting.

て検索結果が変化することを確認する．これは，1 章で述べた提案方式の特徴（2）の検証である．具体的には，実験 A と同様のデータセットを使用して，同じ検索語（事象 44）を発行し原因事象検索を行う．まったく重要度（重み）を設定しない場合は，表 2 の原因事象検索のようになる．本実験では，表 2 の原因事象検索において，24 位にランクされている事象 16 に着目し，検索語事象ベクトルの事象 16 に相当する要素に重みを与えたときの順位，および，相関度の変化を調べる．1～10 の 10 種類の重みを用意し，それぞれを検索語事象ベクトルの事象 16 に相当する要素に掛けることによって，計 10 回原因事象検索を行う（重みが“1”のときは実験 A の結果と同様になる）．10 回の検索に対する事象 16 の順位と相関度の変化をグラフとして表現したものをそれぞれ図 6，図 7 に示す．

[考 察]

図 6 より，事象 44 の検索語ベクトルの事象 16 の要素に高い重みを設定するほど，事象 16 が高い順位で検索されていることが分かる．この場合，重みを“7”

以上に設定すると、検索対象である事象 16 が 1 位で検索されるようになっている。また、図 7 より、同様に事象 16 に対する重み高く設定するほど、検索対象である事象 16 が高い相関度を持つ事象として検索されていることが分かる。このことは、検索者の与える検索語事象ベクトルの特定の要素に、自由に重みを設定し、特定の事象の重要度を変化させることが可能であることを示している。

5. 実データへの適用

本章では、提案方式を宇宙工学分野に適用し、本方式の実現可能性について検証する。ここでは、宇宙航空研究開発機構（以下 JAXA）で提供されている不具合情報を対象データとして用いる。これらのデータに提案方式を適用することによって、宇宙工学分野において発生しうる不具合（事象）に関する原因や、その不具合から引き起こる可能性のある結果についての事象群を獲得することが可能となる。

5.1 実験環境

JAXA において提供されている実事象データ・実文書データを対象とした検索システムを構築した。

5.1.1 宇宙工学関連事象データ

JAXA 内で独自に管理された事象データの分類表の中から、「不具合現象の分類」、「不具合原因の分類」に記載されている単語群を本システムで使用する事象群とした。この分類表に示されている単語群は、以下のような特徴を有する。

- 開発・製造から射場整備場までのロケットおよび人工衛星の不具合についての単語が示されている。
- 単語数は、「不具合現象の分類」が 246 語、「不具合原因の分類」が 120 語ある。

これらの分類表に記載されている計 366 個の事象を対象として、2 章に示した方法を用い、JAXA と関係のある宇宙工学の専門家の判断によってそれぞれの事象のベクトル化を行った。

5.1.2 検索対象文書データ

検索対象文書として、JAXA 内で独自に管理されている「不具合報告書」253 件、および、「信頼性技術情報」95 件を使用した。これら計 348 件を対象として、5.1.1 項の事象データ群の中からそれぞれの文書に関係の深いメタデータを付与した。メタデータ付与に関しては、宇宙工学研究分野の専門家が行った。

5.1.3 検索システムの構築

5.1.1, 5.1.2 項で作成した事象データ・文書データそれぞれをテキストデータの形式でデータベースを構築した。ベクトル間の相関度計量に関しては、意味の数

学モデルによる意味的連想検索^{8),9)}を使用した。実験システムを構築した計算機の CPU は Intel Pentium III 730 MHz, メモリは 384 MB, OS は Linux kernel 2.4.7 である。

5.2 実験方法

実験結果を評価するために、下記の 2 つの評価関数 EV_{ca} , EV_{ef} を定義する。本実験では、5.1.3 項で構築した検索システムにおいて、上位にランキングされる文書では、 EV_{ca} および EV_{ef} の値が高いことを確認する。また下位にランキングされる文書ほど、 EV_{ca} および EV_{ef} の値が低いことを確認する。

EV_{ca} は、原因事象検索についての評価式である。 EV_{ca} は、事象単語 w を用いて検索したときに、文書 i に設定されているメタデータ群 mds_i 中に「検索語自身 w 」および「 w の原因である事象単語群」が含まれている場合に高い値をとるように作用し、「検索語自身 w 」および「 w の原因である事象単語群」以外の事象単語が含まれている場合に低い値をとるように作用する。また、 EV_{ef} は、結果事象検索についての評価式である。 EV_{ef} は、事象単語 w を用いて検索したときに、文書 i のメタデータ群 mds_i 中に「検索語自身 w 」および「 w の結果である事象単語群」が含まれている場合に高い値をとるように作用し、「検索語自身 w 」および「 w の結果である事象単語群」以外の事象単語が含まれている場合に低い値をとるように作用する。

評価式 1 原因事象検索についての評価式

$$EV_{ca}(w, mds_i) = p_0 \cdot N_0(w, mds_i) + \sum_{k=1}^n p_k \cdot N_{ca(k)}(w, mds_i) - q \cdot N_{ca(OTHER)}(w, mds_i)$$

評価式 2 結果事象検索についての評価式

$$EV_{ef}(w, mds_i) = p_0 \cdot N_0(w, mds_i) + \sum_{k=1}^n p_k \cdot N_{ef(k)}(w, mds_i) - q \cdot N_{ef(OTHER)}(w, mds_i)$$

ここで、

$p_k > 0, q > 0$: 重み係数

w : 検索語として入力する事象単語

mds_i : 事象単語 w で検索した結果、 i 位にランキングされた文書に設定されているメタデータ群

$N_0(w, mds_i)$: 事象単語 w に対し、 w 自身が mds_i に含まれている個数

$N_{ca(k)}(w, mds_i)$: 事象単語 w に対し, w の k 個前の原因である事象単語が mds_i に含まれている個数

$N_{ca(0)}(w, mds_i)$: 事象単語 w に対し, mds_i の個数 $-N_0(w, mds_i)$
 $-\sum_{k=1}^n N_{ca(k)}(w, mds_i)$

$N_{ef(k)}(w, mds_i)$: 事象単語 w に対し, w の k 個先の結果である事象単語が mds_i に含まれている個数

$N_{ef(0)}(w, mds_i)$: 事象単語 w に対し, mds_i の個数 $-N_0(w, mds_i)$
 $-\sum_{k=1}^n N_{ef(k)}(w, mds_i)$

本実験では, 提案方式により構築した検索システムにおいて, 原因事象検索, 結果事象検索それぞれの検索について, 以下の事象単語を検索語として入力した場合の検索結果を示す. なお, 本実験では, 評価式に対して, $p_0 = p_1 = p_2 = 1, p_n = 0 (n > 2), q = 1/2$ として実験を行った. これらのパラメータ設定を行った理由としては, 最も単純な値であり, それに対する評価式が算出する結果が解釈しやすいと判断したからである.

(a) 原因事象検索に用いた検索語:

- w_1 : 「孔位置/ピッチ不良」
- w_2 : 「色むら・光沢むら・色違い・しみ・汚れ・変色・斑点」
- w_3 : 「ハンダ付不良」
- w_4 : 「輸入品」
- w_5 : 「耐環境性能検討不十分」
- w_6 : 「出力低下・出力増大」
- w_7 : 「絶縁不良」

(b) 結果事象検索に用いた検索語:

- w_8 : 「ハンダ付不良」
- w_9 : 「異部品装着」
- w_{10} : 「計測器精度不良」
- w_{11} : 「図面指示誤り, 漏れまたは不明確」
- w_{12} : 「回路設計不十分」
- w_{13} : 「耐環境性能検討不十分」
- w_{14} : 「絶縁不良」

5.3 実験結果

実験結果を図 8, 図 9 に示す. 検索語 $w_1 \sim w_7$ についての結果は, 図 8 に示されている. また, 検索語 $w_8 \sim w_{14}$ についての結果は, 図 9 に示されている.

ここで, 図中に示されている EV_{ca_avr} , および EV_{ef_avr} は, 順位が 1 位から rank 位までの文書に対する EV_{ca} または EV_{ef} の平均値である. すなわち,

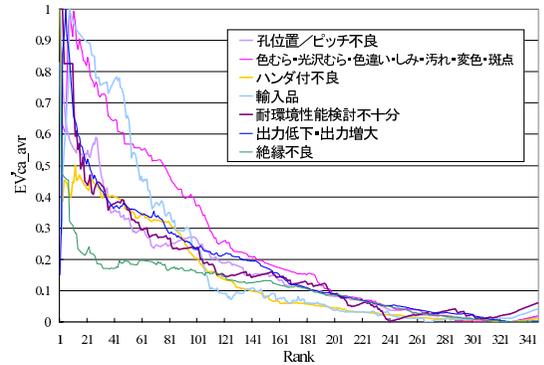


図 8 実事象データを対象とした原因事象検索

Fig. 8 Cause search results for a real event data set.

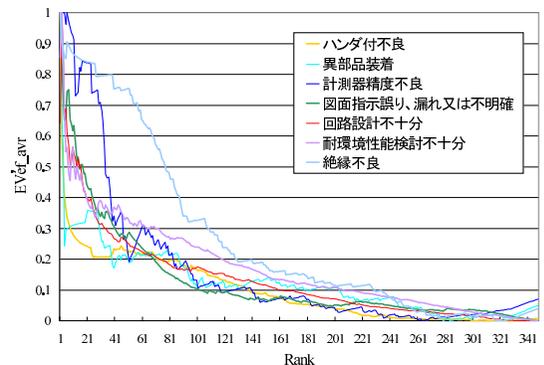


図 9 実事象データを対象とした結果事象検索

Fig. 9 Effect search results for a real event data set.

$$EV_{ca_avr}(w, rank) = \frac{\sum_{i=1}^{rank} EV_{ca}(w, mds_i)}{rank}$$

$$EV_{ef_avr}(w, rank) = \frac{\sum_{i=1}^{rank} EV_{ef}(w, mds_i)}{rank}$$

また, EV'_{ca_avr} および EV'_{ef_avr} は, EV_{ca_avr} および EV_{ef_avr} が, 0 から 1 の範囲になるように正規化した値である.

$$EV'_{ca_avr} = \frac{EV_{ca_avr} - EV_{ca_min}}{EV_{ca_max} - EV_{ca_min}}$$

$$EV'_{ef_avr} = \frac{EV_{ef_avr} - EV_{ef_min}}{EV_{ef_max} - EV_{pheno_min}}$$

ここで,

$EV_{ca_max}(w)$: 事象単語 w で検索したときの EV_{ca} の最大値

$EV_{ca_min}(w)$: 事象単語 w で検索したときの EV_{ca} の最小値

$EV_{ef_max}(w)$: 事象単語 w で検索したときの EV_{ef} の最大値

$EV_{ef_min}(w)$: 事象単語 w で検索したときの EV_{ef} の最小値

5.4 考 察

実験結果より, $w_1 \sim w_7$ の事象単語をシステムに入力し, 原因事象検索した場合について, 上位に検索される文書は EV'_{ca} の値が高く, 下位に検索される文書ほど EV'_{ca} の値が低くなる傾向にあることを確認できる. この結果は, 原因についての検索を行った場合に, ある事象 w の「原因である事象」について記述された文書群が上位に検索され, 原因とは関係のない事象について記述された文書群が下位に検索されることを示している. また, $w_8 \sim w_{14}$ の事象単語をシステムに入力し, 結果事象検索した場合について, 上位に検索される文書は, EV'_{ef} の値が高く, 下位に検索される文書ほど, EV'_{ef} の値が低くなる傾向にあることを確認できる. この実験結果は, 結果についての検索を行った場合に, ある事象 w の「結果である事象」について記述された文書群が上位に検索され, 結果とは関係のない事象について記述されている文書群が下位に検索されることを示している.

以上の実験結果により, 原因事象検索では, ある事象 w の「原因である事象」について記述された文書群が上位に検索され, 原因とは関係のない事象について記述された文書群が下位に検索されることが, 検索結果全体として確認できた. また, 結果事象検索では, ある事象 w の「結果である事象」について記述された文書群が上位に検索され, 結果とは関係のない事象について記述された文書群が下位に検索されることが, 検索結果全体として確認できた.

検索精度評価を行うときは, 適合率・再現率²⁾を使用するのが一般的である. 適合率・再現率を使用するには, あらかじめ検索語に対する正解セットを用意する必要がある. 提案方式を用いた場合, 検索対象領域の専門家が検索語事象データに対しての正解文書セットを用意し, 評価実験を行うことが考えられる. しかし, 専門家が検索語事象データに対して, 間接的な因果関係まで考慮した正解文書セットを設定するのは非常に困難であると考えられる. このような理由から, 本実験では, 情報検索の分野で一般的に使用される評価方法を適用するのではなく, 独自に考案した評価式 (EV_{ca} , EV_{ef}) の用いた評価を行い, 提案方式を用いて因果関係が計量可能であることを評価した.

実験結果 (図 8, 9) の中で, グラフが上下する部分については次のように考察できる. 評価式 (EV_{ca} , EV_{ef}) の評価方法は, 単純に直接的な因果関係ほど高い値を算出するようになっているのに対して, 提案方式を用いた因果関連性計量では, 4.2 節の考察で述べたように, 間接的な因果関係を持つ事象であっても

多くの直接的な事象の因果となっていれば高い相関度を算出するようになっている. この両者の違いが, 図中のグラフのぶれとなって現れていると考えられる. しかし, 全体として右下がりの傾向を持ったグラフになっていることが明確に示されており, 提案方式の有効性を評価するには十分であると考えられる.

6. 実験のまとめ

ここでは, 本研究における提案方式, 行った実験に関する時間的コストに関する考察を示す. ここでのコストとは, 対象領域における事象群のベクトル化に関するコストと, 検索者の検索時に発生する事象ベクトル間の相関度計量に関するコストのことを意味しており, それぞれについての考察を示す.

6.1 対象領域における事象群のベクトル化に関するコスト

提案方式を用いて検索を行うための前段階として, 検索対象領域における事象群をそれぞれベクトル化する必要がある. 本来, 事象間の因果関係は, 対象とする領域の専門家が知識として有している. 提案方式における事象群のベクトル化に関しては, 2 章で述べたとおり, 対象領域の専門家の判断によって直接的な因果関係の有無を記述することだけを必要としている (間接的な因果関係を記述する必要はなく, さらに, 事象間の因果関係の確率などを明確に設定する必要もない). このことは, 対象領域の専門家の事象間因果関係に関する知識を検索可能な形式に表現することに關して, 専門家に要する人的コストを最小限におさえることが可能であることを示している. 実際に, 5 章で示した, JAXA における宇宙工学分野に関する実事象データ群 366 件のベクトル化に要した時間は, JAXA 内の 3 名の宇宙工学の専門家が, それぞれ約 30 時間である. この時間は, 専門家の因果関連性に関する知識が検索可能となることを考慮すると, 現実的に作業可能な時間であると考えられる.

より多くの事象群が発生する検索対象領域について, 事象群のベクトル化に要するコストに関しては次のように考察できる. 検索対象領域の事象群が多い場合は, 事象間の因果関係がより複雑になり, 事象群のベクトル化に多大な時間を要する可能性がある. このような場合には, 検索対象領域をさらに小規模な複数の領域に分割し, 分割した小領域ごとに事象群のベクトル化を行う. さらに, これらの複数のベクトル群に対して, 文献 6), 12) において提案されている検索空間統合方式を適用することによって, 複数の小領域の事象ベクトル群を統合することが可能となり, 全事象を対象と

した因果関係検索が可能となると考えられる。このことは、提案方式における事象群のベクトル化に関する作業時間の拡張可能性を示している。実際に提案方式に対して検索空間統合方式を適用した実験に関しては、今後の研究課題とする。

6.2 検索者の検索時に発生する事象ベクトル間の相関度計量に関するコスト

提案方式を用いて検索を行ったときに発生する事象ベクトル間の相関度計量の際に関するコストは、従来のベクトル空間モデルによる検索方式を用いたときのコストとほぼ同じであるといえる。ベクトル間の相関度計量に関するコストは、検索対象である事象群の数に依存する。つまり、事象群の数が多ければ事象ベクトルの要素の数が増え、相関度計量(内積やコサイン)に要する時間は増大する。しかし、事象 A・B 間の相関度計量と事象 A・C 間の相関度計量はお互いに干渉しない、つまり、複数の相関度計量はそれぞれ独立した計算が可能であるので、それぞれの相関度計量を複数のプロセッサに割り当てるのが可能である。すなわち、提案方式における事象ベクトル間の相関度計量は、プロセッサの数に拡張可能であるといえる。

また、検索対象事象ベクトル群からなる行列は「0」の多いスパース(sparse)である場合には、行列の圧縮形式⁵⁾を適用できる可能性がある。5章で示した実験で用いた JAXA の実事象データの例では、各事象データには平均 15 個の要素(事象)に値が付与されている。つまり、残りの 251 の要素は「0」が付与されており、全体としてスパースな行列を形成している。このスパースの行列に対して行列の圧縮形式を適用することによって、事象ベクトル群を計算機上に効率的に格納でき、さらに、事象ベクトル間の相関度計量も効率化できると期待できる。このことについては今後の課題とする。

7. 結 論

本稿では、事象間の因果関連性計量をともなったベクトル空間検索方式を提案した。提案方式は、さらに、検索対象領域における事象間の因果関連性が計量可能なベクトル空間を生成する方式と、ベクトル空間上での検索方式に分割することができ、それぞれの方式について述べた。1章で述べたように、提案方式の特徴は次の3つにまとめられる。

- (1) ベクトルデータに対する様々な計算手法を適用可能
- (2) 特定の事象データを重要視した検索の実現
- (3) 因果関連性計量をともなった文書検索の実現

これらの特徴を検証するために、擬似事象データ・文書データ、および、宇宙工学分野の実データを用いた実験を行い、提案方式の有効性、および、実現可能性について示した。

今後は、提案方式を実事象データ・文書データに適用し、因果関連性計量をともなった検索エンジンの構築を目指す。さらに、グラフ構造でたどれるような因果関係だけでなく、暗黙的な因果関係が計量できるように本方式を拡張する予定である。また、A B C A というように、事象間の因果関係が循環するような領域にも適用可能な方式も導入していく予定である。

参 考 文 献

- 1) 甘利俊一, 佐藤俊哉, 竹内 啓, 狩野 裕, 松山 裕, 石黒真木夫: 多変量解析の展開—隠れた構造と因果を推理する, 岩波書店(2002).
- 2) Baeza-Yates, R., Ribeiro-Neto, B.: *Modern Information Retrieval*, Addison Wesley (1999).
- 3) Bollen, K.A.: *Structural Equations With Latent Variables*, John Wiley & Sons Inc (1989).
- 4) Deerwester, S.C., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R.A.: Indexing by latent semantic analysis, *J. Am. Soc. Inf. Sci.*, Vol.41, No.6, pp.391-407 (1991).
- 5) Duff, I., Grimes, R.G. and Lewis, J.G.: Users' guide for the Harwell-Boeing sparse matrix collection (Release I), Technical Report TR/PA/92/86, CERFACS (1992).
- 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.SIG 5 (TOD14), pp.37-53 (2002).
- 7) Jain, A.K., Murty, M.N. and Flynn, P.J.: Data Clustering: A Review, *ACM Computing Surveys*, Vol.31, No.3 (1999).
- 8) 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*, Sheth, A. and Klas, W. (Eds), Chapter 7, McGrawHill (1998).
- 9) 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).
- 10) Pearl, J.: *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Francisco, CA (1988).
- 11) Salton, G., Wong, A. and Yang, C.S.: A vector space model for automatic indexing, *Comm.*

ACM, Vol.18, No.11, pp.613-620 (1975).

- 12) 鷹野孝典, 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プロセスの実現, DBSJ Letters, Vol.1, No.1, pp.55-58 (2002).
- 13) 鷹野孝典, 冨子泰三, 清木 康, 但田育直, 波内みさ: 時間的因果関係を扱う動的な文脈解釈を伴った意味的連想検索方式の実現, 情報処理学会研究報告 DBS-131, pp.483-489 (2003).
- 14) 冨子泰三, 吉田尚史, 清木 康: ドキュメントデータ群を対象とした文脈依存動的クラスタリングの再帰的適用による意味的知識発見方式, 情報処理学会論文誌: データベース, Vol.43, No.SIG 2 (TOD13), pp.216-230 (2002).
- 15) 冨子泰三, 鷹野孝典, 清木 康: 事象データ群の時間因果関係を扱う意味的連想検索方式, 情報処理学会研究報告 DBS-130, pp.71-77 (2003).

(平成 15 年 12 月 20 日受付)

(平成 16 年 4 月 6 日採録)

(担当編集委員 岸田 和明)



冨子 泰三 (学生会員)

1999 年慶應義塾大学環境情報学部卒業. 2002 年同大学院政策・メディア研究科修士課程修了. 現在, 同大学院同研究科博士課程に在学中. データベースシステム, データマイニング,

情報検索に関する研究に従事. 日本データベース学会会員.



清木 康 (正会員)

1978 年慶應義塾大学工学部電気工学科卒業. 1983 年同大学院工学研究科博士課程修了. 工学博士. 同年日本電信電話公社武蔵野電気通信研究所入所. 1984 年~1995 年筑波

大学電子・情報工学系講師, 助教授を経て, 1996 年慶應義塾大学環境情報学部助教授, 1998 年同学部教授. データベースシステム, 知識ベースシステム, マルチメディアシステムの研究に従事. ACM, IEEE, 電子情報通信学会, 日本データベース学会各会員.



鷹野 孝典 (学生会員)

1998 年慶應義塾大学環境情報学部卒業. 同年 (株) 図研入社. 2003 年慶應義塾大学大学院政策・メディア研究科修士課程修了. 同年より, 同大学院同研究科博士課程に在学. データベースシステムの研究に従事. ACM, 日本データベース学会各会員.



波内 みさ (正会員)

1988 年筑波大学第三学群情報学類卒業. 1990 年筑波大学理工学研究科修了. 同年日本電気株式会社に入社し, オブジェクト指向データベース管理システムの研究に従事. 2000 年宇宙開発事業団 (現宇宙航空研究開発機構) 入社.



但田 育直

2000 年京都工芸繊維大学工芸学部電子情報工学科卒業. 2002 年東京工業大学大学院総合理工学研究科知能システム科学専攻修士課程修了. 同年より宇宙開発事業団 (現宇宙航空研究開発機構) に勤務.