社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－DBS－130 (9)
2003－ＦＩ－ 71 (9)
2003／5／22

事象データ群の時間的因果関係を扱う意味的連想検索方式

図子泰三[†] 鷹野孝典[†] 清木 康[††]

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，事象間の時間的因果関係を扱う意味的連想検索方式を示す．本方式の実現により，検索者の与える事象に時間的因果関係のある事象について記述された文書データを検索することが可能となる．擬似的な事象データ・文書データを用いた実験により，本方式の有効性を明らかにする．

A Semantic Associative Search Method with Temporal Cause-and-effect Relationship for Event Data Sets

Taizo Zushi,[†] Kosuke Takano[†] and Yasushi Kiyoki[††]

In this paper we discuss a semantic associative search method with temporal cause-and-effect relationship between events. The method is used for the domain in which a query given by a user can be regarded as an event. The method enables to retrieve a document data set having the relationship with a query. We clarify effectiveness of the method by an experiment using a dummy event and document data set.

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と確認されている．

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が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）と捉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，事象間の時間的因果関係が計量可能な系が必要となる．

本稿では，検索対象領域における事象間の時間的因果関係が計量可能なベクトル空間検索方式を示す．本方式は，意味の数学モデル [3),4)] による意味的連想検索方式を対象として，適用する方式である．意味の数学モデルでは，検索対象領域における基本単語群を特定し，各単語の定義をベクトルデータとして表現することによって，単語間の意味的な類似性を検索者の文脈に応じて計量することが可能な意味空間を生成する．本稿に示す方式では，検索対象領域において起こり得る事象群を特定し，各事象間の時間的因果関係に応じてベクトルデータを作成することによって，事象間の時間的因果関係が計量可能な意味空間を生成することが可能となる．

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 [5)] や Latent Sematic Indexing(LSI)[2)] 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もつベクトルで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時間的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，検索者の検索語と時間的因果関係のある文書データを検索することはできない．

意味の数学モデルでは，検索空間を生成する際に，図１に示すような意味空間生成用メタデータ（行列）

† 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

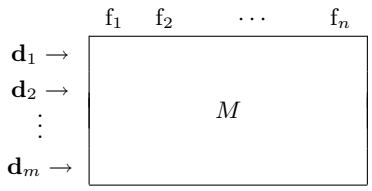

図 1 データ行列 M によるメタデータの表現

が作成される．意味空間生成用メタデータでは，検索対象領域における特徴語群を特定し，その特徴語間の意味づけを行列によって表現することによって，検索対象メディアデータが写像可能な意味空間を生成することができる．

本稿では，意味の数学モデルにおける意味空間生成用メタデータを構成する特徴語群を，事象データ群に代替し，その事象データ間の時間的因果関係に応じて行列を作成する方式について述べる．さらに，検索者の検索目的に応じた，文書データに付与するメタデータ，および，検索者の検索語のベクトル生成法について述べる．

2. 意味の数学モデルの概要

各分野における基本用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，意味の数学モデルによるメディアデータ検索方式の概要を示す．

(1) **メタデータ空間 $\mathcal{MDS}$ の設定**

m 個の基本データについて各々n 個の特徴（f_1,f_2,$\cdots$,f_n）を列挙した特徴付ベクトル $\mathbf{d}_i(i = 1,\cdots,m)$ が与えられているものとし，そのベクトルを並べて構成する $m \times n$ 行列を M とおく（図 1）．行列 M より，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 $\mathcal{MDS}$) を設定する．

(2) **メディアデータのメタデータをメタデータ空間 $\mathcal{MDS}$ へ写像**

設定されたメタデータ空間 $\mathcal{MDS}$ 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，検索対象データのメタ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意味的な関係を空間上での距離として計算することが可能となる．

メディアデータ P には，メタデータとして t 個の基本データ $\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_t$ が以下のように付与されていることを前提としている．

$$P = \{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_t\}. \qquad (1)$$

各基本データは，ベクトル表現された特徴を持っている．

$$\mathbf{w}_i = (f_{i1}, f_{i2}, \cdots, f_{in}). \qquad (2)$$

各メディアデータは，メタデータとして付与されている t 個の基本データが合成されベクトル表現された後，メタデータ空間 $\mathcal{MDS}$ へ写像される．

(3) **メタデータ空間 $\mathcal{MDS}$ 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キストと呼ぶ．このコンテキストを用いてメタデータ空間 $\mathcal{MDS}$ に各コンテキ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 $\mathcal{MDS}$ 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．

この操作により，検索者が与えたコンテキストに対して相関の強い軸のみによる部分空間が選択される．与えられたコンテキストによりダイナミックに選択されたこの部分空間上においてメディアデータベクトルのノルムを計量することにより，与えられたコンテキストに対して意味的に相関の強い検索対象データを，ダイナミックに解釈することが可能となる．

この部分空間選択機構により，各検索対象データについて，与えられたコンテキストを構成する単語群が共通にもつ要素に対応する部分にのみ着目した相関量を計量することが可能となる．すなわち，コンテキストとして与えられる単語群が共通にもつ要素群（軸群）による部分空間を抽出することにより，検索者の意図をシャープに反映した相関量計算が可能になる．

(4) **メタデータ空間 $\mathcal{MDS}$ の部分空間 (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 $\mathcal{MDS}$ 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，与えられたコンテキ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．

また，メディアデータを特徴づける特徴の数

が多い場合，どのような意味空間が選ばれても，意味空間におけるメディアデータのノルムが大きくなる傾向がある．そのため，本来，文脈との相関が強いと考えられるメディアデータベクトルのノルムよりも，特徴の数が多いメディアデータベクトルのノルムが大きくなってしまい，適切な抽出が行われないことがある．そのため，メタデータ空間でのメディアデータベクトルを 2 ノルムで正規化している．

2.1 メタデータ空間生成方式

以下に，メタデータ空間の生成プロセスを示す．

(a) 対象とする分野を表現するために必要な特徴語（以下，feature）群を準備する．対象分野の専門辞書等を用いて，各見出し語を説明している説明文中の単語を抽出し，この集合を feature 群とする．これにより，その分野の意味を表現するのに必要な単語群が定義される．

(b) 対象とする分野の基本的な用語である，基本データ群を準備する．(a) と同様に，専門辞書を用いて，見出し用語群を抽出し，この集合を基本データ群と定義する．

(c) feature 群を用いて，各基本データの特徴付けを行う．同様の専門辞書を用いて基本データの説明文を調べ，説明文をもとに，関係のある feature には 1 を，逆の意味で用いられている feature には-1 を，関係のない feature には 0 を，それぞれ設定する．この方法で，すべての基本データに対して，feature による特徴付けを行う．

(d) 以上の feature による基本データの特徴付けマトリクスから，意味的連想検索のためのメタデータ空間を生成する．

以上のプロセスにより，対象分野における意味の形式的な計算を可能とするメタデータ空間を生成する．

3. 提案方式の概要

本節では，本稿で提案する事象データ群の時間的因果関係を扱う意味的連想検索方式について述べる．2 節で概要を示した意味の数学モデルの中で使用するベクトルデータに関して，事象データ間の時間的因果関係に応じたベクトルデータの作成方法について論ずる．図 2 に示すような擬似的な事象データ（事象 A〜F）を使用して解説する．図 2 における矢印は時間的因果関係を示している．例えば,「事象 B の前には事象 A が発生する（事象 A は事象 B が起こる原因となっている）」,「事象 D が起こると，それに伴って事象 E が起こる（事象 D の結果，事象 E が起こる）」ということを示している．

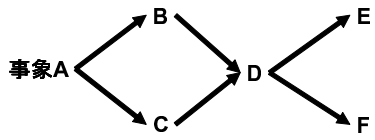

図 2 擬似的な事象データ

3.1 空間生成のためのメタデータ

空間生成のためのメタデータ，すなわち，2 節における行列 M の作成方法について解説する．提案方式では，基本データ，特徴，両方ともに同様の事象データ群を設定するため $m \times m$ の正方行列を形成する．図 2 の例を用いると，事象 A〜F について，6×6 の行列を作成することになる．それぞれの事象ベクトルの要素については，その事象自身と，その事象と直接関係のある事象に「1」を設定し，それ以外の事象には「0」を設定する．例えば，事象 B のベクトルを作成する場合，事象 B 自身と，事象 B と直接関係のある事象 A と事象 D に「1」を設定し，それ以外の事象には「0」を設定する．この作業を全事象ベクトルに適用すると図 3 のようになる．

	A	B	C	D	E	F
A	1	1	1	0	0	0
B	1	1	0	1	0	0
C	1	0	1	1	0	0
D	0	1	1	1	1	1
E	0	0	0	1	1	0
F	0	0	0	1	0	1

図 3 擬似的な事象データの空間生成用メタデータ表現 (M)

3.2 キーワードのためのメタデータ

キーワードのためのメタデータ，すなわち，検索者が検索語として与える事象データのベクトル表現の方法について解説する．キーワードのためのメタデータも空間生成用メタデータと同様に，$m \times m$ の正方行列となる．次に示すように，検索者の検索目的に応じて，異なる二種類の行列データを用意する．3.1 節と同様に，図 3 の例を使用して解説する．

(1) **検索語として与えられる事象の原因となる事象を検索する場合**

事象ベクトルについて，その事象自身と，その事象の直接の原因となっている事象に「1」を

設定し，それ以外に「0」を設定する．図 2 の事象 D を例とすると，事象 D 自身と，事象 D の直接の原因となっている事象 B，C に「1」を設定し，その他の事象には「0」を設定する．同様の作業を全事象について行うと，図 4 のようになる．この行列を M_c とする．

（2） **検索語として与えられる事象によって引き起こる事象を検索する場合**

事象ベクトルについて，その事象自身と，その事象が直接の原因となっている事象に「1」を設定し，それ以外に「0」を設定する．図 2 の事象 D を例とすると，事象 D 自身と，事象 D が直接の原因となっている事象 E，F に「1」を設定し，その他の事象には「0」を設定する．同様の作業を全事象について行うと，図 5 のようになる．この行列を M_r とする．

	A	B	C	D	E	F
A	1	0	0	0	0	0
B	1	1	0	0	0	0
C	1	0	1	0	0	0
D	0	1	1	1	0	0
E	0	0	0	1	1	0
F	0	0	0	1	0	1

図 4 検索事象の原因を検索する場合のキーワードメタデータ (M_c)

	A	B	C	D	E	F
A	1	1	1	0	0	0
B	0	1	0	1	0	0
C	0	0	1	1	0	0
D	0	0	0	1	1	1
E	0	0	0	0	1	0
F	0	0	0	0	0	1

図 5 検索事象によって引き起こる事象を検索する場合のキーワードメタデータ (M_r)

3.3 検索対象文書のためのメタデータ

検索対象文書のためのメタデータの設定方法について解説する．ここでは，一つの文書データに対して，複数個の単語（事象）がメタデータとして付与されているものとする．文書データに付与されているそれぞれの単語はベクトルデータとして表現されており，一つの文書データは，複数のベクトルデータの合成ベクトルとして表現される．キーワードのためのメタデータと同様に，検索者の検索目的に応じて，異なる二種類の行列データを用意する．図 3 の例を使用して，一つの文書データに事象 B，D がメタデータとして付与されている場合のベクトルデータの作成される過程を説明する．

（1） **検索語として与えられる事象の原因となる事象を検索する場合**

事象ベクトルとして，3.2 節における図 5 の行列に示されている各ベクトル，すなわち，M_r を使用する．文書データに付与されている事象（B，D）ベクトルの合成ベクトルをこの文書の文書ベクトルとする．つまり，図 6 の上部のようになる．

（2） **検索語として与えられる事象によって引き起こる事象を検索する場合**

事象ベクトルとして，3.2 節における図 4 の行列に示されている各ベクトル，すなわち，M_c を使用する．同様の作業を行うことによって，図 6 の下部のようになる．

各文書のメタデータのベクトルを合成する方法としては，**??**節に記述したとおりである．

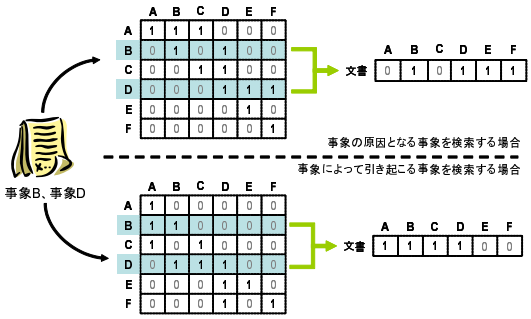

図 6 検索対象用メタデータの作成

3.4 各メタデータの組み合わせ

実際に，検索者がある事象を検索語として与えて検索を行う場合，検索目的に応じて，以下の 2 通りの行列データの組み合わせが考えられる．

（1） **検索語として与えられる事象の原因となる事象について記述されている文書を検索する場合**

空間生成用メタデータとして行列 M，キーワードのためのメタデータとして行列 M_c，検索対象のためのめたデータとして M_r を使用する．

（2） **検索語として与えられる事象によって引き起こる事象について記述されている文書を検索する**

場合

空間生成用メタデータとして行列 M，キーワードのためのメタデータとして行列 M_r，検索対象のためのめたデータとして M_c を使用する．

4. 実　　験

本節では，擬似的な文書データ群を対象とした実験を行うことによって，提案方式である事象データ群の時間的因果関係を扱う意味的連想検索方式の有効性について検証する．

4.1 実験環境

実験を行うための擬似的な検索対象領域における事象群を図 7 のように設定した．図 7 における矢印は，図 2 と同様に，事象間の時間的因果関係を示している．例えば，事象 p の起こる原因として事象 β が存在し，事象 p の結果として事象 α が起こることを示している．これらの事象群の時間的因果関係を 3 節において示した方法を用いて 3 種類の 9×9 の行列 M，M_c，M_r を作成した．

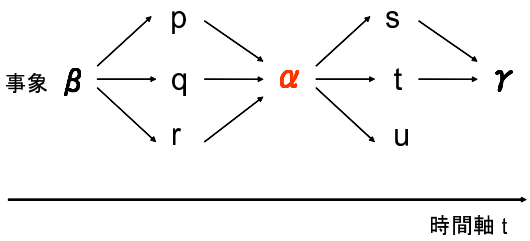

図 7　実験のための擬似的な検索対象領域における事象群

また，検索対象として表 1 に示すような 17 件の擬似的な文書群を作成した．例えば，doc01c の文書には，事象 p，事象 q，事象 r に関する内容が記述されていることを意味している．文書 ID の末尾に「c」が付与されている文書（例えば doc01c）は，事象 α の原因として起こる事象に関する内容が記述されていることを示す．同様に，文書 ID の末尾に「r」が付与されている文書（例えば doc09r）は，事象 α によって引き起こる事象に関する内容が記述されていることを示す．doc17n は事象 α に関する内容が記述されていることを示す．

4.2 実験方法

4.1 節で構築した実験環境を対象として，2 節で示した意味的連想検索方式を適用することによって，検索者の検索語として与える事象と時間的因果関連性の高い文書データが検索されることを確認する．検索語として「事象 α」を与え，以下に示す 3 種類の目的に応じた行列の組み合わせを用いた検索を行った．

表 1　検索対象用擬似文書データ

文書 ID	メタデータ
doc01c	p q r
doc02c	p q
doc03c	p r
doc04c	q r
doc05c	p
doc06c	q
doc07c	r
doc08c	β
doc09r	s t u
doc10r	s t
doc11r	s u
doc12r	t u
doc13r	s
doc14r	t
doc15r	u
doc16r	γ
doc17n	α

(1) **事象 α の原因となる事象について記述されている文書の検索**

空間生成用メタデータとして行列 M，キーワードのためのメタデータとして行列 M_c，検索対象文書のためのメタデータとして行列 M_r を使用する．

(2) **事象 α の後に引き起こる事象について記述されている文書の検索**

空間生成用メタデータとして行列 M，キーワードのためのメタデータとして行列 M_r，検索対象文書のためのメタデータとして行列 M_c を使用する．

(3) **事象 α に関連する内容について記述されている文書の検索**

空間生成用メタデータ，キーワードのためのメタデータ，検索対象文書のためのメタデータともに行列 M を使用する．

4.3 実験結果

4.2 節の検索 (1)，(2)，(3) に対する結果をそれぞれ表 2, 3, 4 に示す．

4.4 考　　察

検索 (1) に対する結果（表 2）より，事象 α の原因となる事象について記述された文書（文書 ID の末尾が「c」である文書）が高い相関度で上位に検索されていることが確認できる．検索 (2) に対する結果（表 3）より，事象 α の後で引き起こる事象について記述された文書（文書 ID の末尾が「r」である文書）が高い相関度で上位に検索されていることが確認できる．検索 (3) に対する結果（表 4）においては，事象 α 自身について記述された文書（doc17n）が高い相関度で

表 2 検索 (1) に対する結果

順位	文書 ID	相関値
1	doc01c	0.880656
2	doc03c	0.700678
3	doc02c	0.700678
4	doc04c	0.700678
5	doc08c	0.687319
6	doc05c	0.597682
7	doc07c	0.597682
8	doc06c	0.597682
9	doc16r	0.429062
10	doc12r	0.132624
11	doc11r	0.132624
12	doc14r	0.103519
13	doc13r	0.103519
14	doc17n	0.102757
15	doc15r	0.069236
16	doc09r	0
17	doc10r	0

表 3 検索 (2) に対する結果

順位	文書 ID	相関値
1	doc16r	0.911037
2	doc09r	0.847865
3	doc10r	0.800973
4	doc15r	0.720557
5	doc11r	0.669647
6	doc12r	0.669647
7	doc13r	0.544086
8	doc14r	0.544086
9	doc08c	0.353479
10	doc07c	0.118614
11	doc06c	0.118614
12	doc05c	0.118614
13	doc04c	0
14	doc01c	0
15	doc02c	0
16	doc03c	0
17	doc17n	0

表 4 検索 (3) に対する結果

順位	文書 ID	相関値
1	doc17n	0.888127
2	doc15r	0.403939
3	doc01c	0.157891
4	doc10r	0.125146
5	doc13r	0.118606
6	doc14r	0.118606
7	doc09r	0.107018
8	doc04c	0.104167
9	doc03c	0.104167
10	doc02c	0.104167
11	doc12r	0.074108
12	doc11r	0.074108
13	doc07c	0.060551
14	doc06c	0.060551
15	doc05c	0.060551
16	doc08c	0
17	doc16r	0

1 位に検索されていることがわかる．これらの実験結果は，提案方式の実現によって，検索目的に応じた事象間の時間的因果関係を伴った文書検索が可能となることを示している．

5. 結　　論

本稿では，検索対象領域で起こり得る事象間の時間的因果関係を扱う意味的連想検索方式を示した．本方式は，事象間の時間的因果関係をベクトル空間の形式で表現することが可能となることが特徴である．また，検索語として与えられる事象の前を検索する場合と後を検索する場合において，それらに応じた異なる 2 通りのベクトルセットを作成することによって，検索者の目的に応じた文書検索が可能となる．擬似的な事象データ・文書データを対象とした検索実験により，本方式の有効性を確認した．

今後は，大量の実事象データ・文書データを対象とした実験システムを構築し，本方式の有効性評価実験，各ベクトルデータの重み付けの導入，本方式のシーケンシャルパターンマイニング（Sequential Pattern Mining）[1] への適用を行っていく予定である．

参 考 文 献

1) Agrawal, R. and Srikant, R.: “Mining sequential patterns,” International Conference on Data Engineering(ICDE’95), pp.3-14, 1995.
2) Deerwester, S. C., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. A.: “Indexing by latent semantic analysis”, Journal of the American Society for Information Science, Vol. 41, No. 6, pp. 391-407, 1991.
3) Kiyoki, Y., Kitagawa, T. and Hayama, T.: “A Metadatabase System for Semantic Image Search by a Mathematical Model of Mea ning,” Multimedia Data Management – using metadata to integrate and apply digital media –, McGrawHill, A. Sheth and W. Klas(editors), Chapter 7, 1998.
4) 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.
5) Salton, G., Wong, A. and Yang, C. S.: “A vector space model for automatic indexing”, Communications of the ACM, Vol. 18, No. 11, pp.613-620, 1975.
6) 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.
7) 吉田 尚史, 図子 泰三, 清木 康, 北川 高嗣: “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および意味的データマイニング方式," 情報処理学会論文誌：データベース, Vol. 41, No. SIG 1 (TOD5), pp.127-139, 2000.
8) 図子 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.