方向性を有する関係を計量する文脈解釈機能をともなった意味的連想検索方式の実現

A Semantic Associative Search Method with a Context Recognition Function for Measuring Directional Relationships

鷹野 孝典♠　　清木 康♦

Kosuke TAKANO　Yasushi KIYOKI

本稿では，上位下位概念，因果関係のような単語間における方向性を有する意味的関係を対象として，それらの方向性を有する関係についてのベクトル空間上での計量を実現する意味的連想検索方式を示す．我々は，これまでに，因果関係を対象とした計量を行うベクトル空間検索方式を提案してきた．本稿では，その方式を因果関係だけではなく，単語間の上位下位概念の関係に関する意味的関係を計量する方式に拡張する．知識表現として広く用いられている上位下位概念を表すツリー構造を対象とし，上位下位概念の関係を計量するベクトル空間を構成する．上位下位概念の関係および因果関係の計量を組み合わせることにより，様々な検索目的に応じた意味的連想検索を実現することが可能になる．本稿では，宇宙工学分野の単語群について上位下位概念の意味的関係を計量する意味的連想検索空間を構築し，検索実験により実現可能性を確認した．

In this paper, we present a semantic associative method for directional relationships, such as relationship of superior subordinate concept among terms and causal relationships. We have proposed a vector space search method for computing causal relationships. We extend this method to compute directional relationships of superior subordinate concepts. By using general knowledge expression with tree structures, we generate vector spaces for computing these directional relationships. This method deals with the combination of the relationships to make it possible to realize semantic associative search, according to various objectives. We have implemented a semantic search system for computing the relationships among technical terms in the aerospace engineering field. We clarify the effectiveness and feasibility of our system by several experiments.

1. はじめに

近年，様々な組織において，大量の文書データ群がデータベースシステム上で管理され，組織内の共有資源として活用されている．我々は，これまでに，事象間の因果関係が重要な文書データ群間の関係となる専門分野を対象とし，因果関係を計量するベクトル空間検索方式を提案してきた[5,8]．この方式は，専門分野において使用される単語群を対象とし，各単語を単位とした因果関係を表現するベクトル空間を構成し，その空間における単語間，および，単語と文書データ間の関係を距離として計量する計量系の設定により，それらの単語が表す事象との因果関係を有する文書データ群を検索する方式である．文献[5]においては，因果関係計量を実現するベクトル空間検索方式を意味の数学モデル[3]へ適用する方式が示されており，文脈に応じた因果関係に関する相関計量を実現している．

本稿では，その方式を因果関係だけではなく，単語間の上位下位概念の関係に関する意味的関係(図1)を計量する方式に拡張する．因果関係と共に，単語間の上位下位概念関係に関する方向性を有する関係の計量により，事象間の因果関係，上位下位概念の関係が重要な文書データ群間の関係となる専門分野において，様々な検索目的に応じた次の意味的連想検索を実現することが可能になる．

(a) 単語間の上位下位概念関係の方向性を有する関係を計量する場合において，単語間の上位方向と下位方向を規定し，各方向のデータをそれぞれ独立に検索できる．
(b) 文脈に応じた動的な方向関係計量を実現できる．

宇宙工学分野において，下位概念の事象間に因果関係がある場合，その事象の上位概念事象間にも因果関係があるとする知見に基づき，専門家の設定した因果関係だけでなく，上位下位関係の知識を組み合わせることで，原因・結果等を究明したい事象に関連する文書群を獲得できる可能性がある．本研究の応用として，このような上位下位概念の関係および因果関係の計量を組み合わせることにより，それぞれ単独の関係では獲得することのできない文書についての検索方式の実現が考えられる．

本稿では，宇宙工学分野の単語群について上位下位概念関係を表すツリー構造による知識表現を対象として適用し，上位下位概念関係を計量する意味的連想検索空間を構築した．検索実験により，「上位概念」，「下位概念」について，それぞれ独立な文書データ検索を実現可能なことを示した．

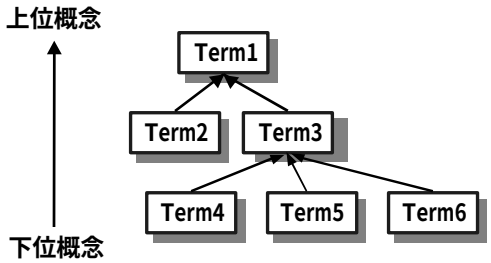

図 1 単語間の上位下位概念についての意味的関係
Fig.1 Superior-subordinate relationships among terms

2. 関連研究

SMART システム[6]や LSI[2]においては，文書データ中に出現する単語の出現頻度に基づいてベクトル表現された検索語と文書データ間において，意味的な等価性や類似性についての静的な関係の計量を行う．また，意味の数学モデルによる意味的連想検索方式[3]では，分野別の専門知識に基づいて，その分野の「意味」を形式的に計量することができるベクトル空間を生成し，そのベクトル空間上において，利用

♠学生会員　慶應義塾大学政策・メディア研究科
kos@sfc.keio.ac.jp
♦正会員　慶應義塾大学環境情報学部
kiyoki@sfc.keio.ac.jp

者が指定した文脈に応じた，意味的に近い情報の動的な検索を実現する．しかし，因果関係や単語間の上位下位関係など，方向を有する関係を計量する場合において，単語間の類似性を算出するための計量系のみでは十分ではない．提案方式においては，方向関係が，単語間や文書間の関係として重要であるデータ群を獲得するための計量系を実現している．

また，提案方式では，単語間の上位下位概念関係における上位方向，下位方向を表現するベクトルを組み合わせ(図 3)，上位関係，下位関係を独立に，それぞれの検索目的に適した相関量計算を直接行うことによる文書検索を実現しており，シソーラスを用いた query expansion[1]のように，まず検索語に関するキーワード展開を行った後，さらにパターンマッチングやベクトル演算による意味的な等価性や類似性の関連性計量により情報獲得を行う方式とは異なる．また，本方式は，単語間の上位下位概念関係のような方向性を有する関係の計量を目的としており，query expansion によるキーワード展開の対象として一般的であるが，類似的な関連性計量を主目的とした関連語[7]は扱わない．

3. 方向性を有する関係を扱うベクトル空間生成方式

単語間の上位下位概念関係の方向性を有する関係を扱うベクトル空間生成方式の概要について述べる．本方式では，単語間の上位下位概念関係を表すツリー構造による知識表現(図 1)を対象として，検索対象領域における単語間の上位下位概念の関係を計量可能なベクトル空間を生成する．このベクトル空間を構成するための特徴行列を，上位下位関係行列と呼ぶ．

3.1 上位下位関係行列の生成

上位下位関係行列として，3つの特徴行列*M*, *Mp*, *Mc*(図 2)を生成するステップを，以下に示す．*M*, *Mp*, *Mc*は，それぞれ，単語の上位下位概念の両方向，上位概念方向，および下位概念方向を表すベクトル群より構成される特徴行列である．

Step-I feature(特徴語) 群，基本データ群の設定

3つの上位下位関係行列*M, Mp, Mc*に対し，n個の単語群*E1*～*En*をfeature，基本データとして，それぞれ上位下位関係行列の横軸，縦軸に設定し，n×nの正方行列を生成する．

Step-II 特徴付けの設定

Step-Iの*M, Mp, Mc*について，特徴付けの設定を行う．

M :基本データ中の単語について，featureの中から，自分自身，上位概念，および下位概念である単語にそれぞれ1を設定し，それ以外の単語に0を設定する．例えば、単語*Eh*の上位概念語が*Ei*，下位概念語が*Ej*，*Ek*であるならば，基本データ単語*Eh*には，*Eh*，*Ei*，*Ej*および*Ek*に相当するfeatureについて1を設定する．この操作を基本データ中の全単語に適用し，*M*を生成する．

Mp:同様に，基本データ中の全単語について，自分自身を表す単語，および上位概念である単語に1を設定し，それ以外の単語に0を設定し，*Mp*を生成する．

Mc:同様に，基本データ中の全単語について，自分自身を表す単語，および下位概念である事象単語に 1 を設定し，それ以外の単語に 0 を設定し，*Mc* を生成する．

3.2 上位下位関係行列の組の設定

3.1 節で生成した*M, Mp, Mc*を用いて，検索目的に応じた上位下位関係行列の組VS−1～VS−3(図3)を設定する．検索対象データを表すベクトル，検索語を表すベクトル（以下、それぞれ検索対象データベクトル，検索語ベクトルと呼ぶ）を形成するための上位下位関係行列として，(A)上位概念である単語群を検索する場合は，それぞれ*Mc, Mp*を設定し，この行列の組をVS−1とする． (B)下位概念である単語群を検索する場合は，それぞれ*Mp*, *Mc*を設定し，この行列の組をVS−2とする． (C)上位概念である単語群，下位概念である単語群の両方について検索する場合は，ともに*M*を設定し，この行列の組をVS−3とする（図3）．VS−1は，検索語ベクトルの形成のために上位概念方向，検索対象データベクトルの形成のために下位概念方向を表現するベクトル群を用いており，単語間の上位概念についての相関量計算を実現する．同様に，VS−2は単語間の下位概念についての相関量計算，VS−3は上位下位の両概念についての相関量計算を実現する．

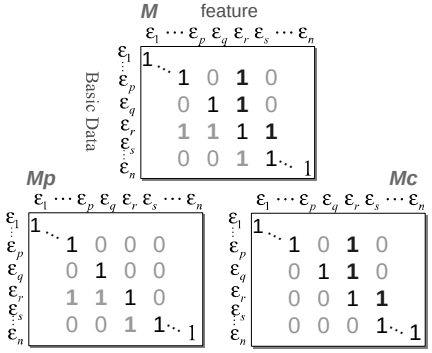

図 2 上位下位関係行列

Fig.2 Matrices with superior-subordinate relationships

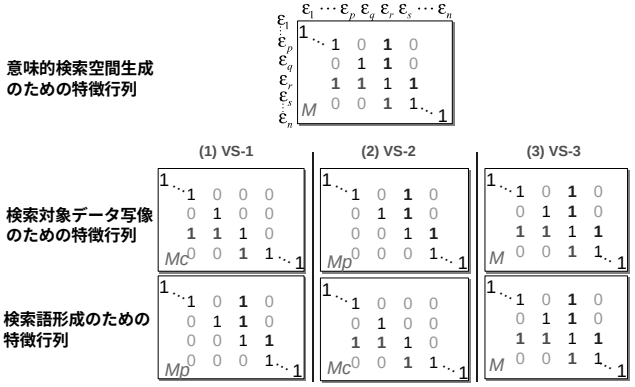

図 3 上位下位関係行列の組の設定

Fig.3 Sets of *superior-subordinate relation* matrices

3.3 意味の数学モデル[3]への適用

上位下位関係行列を，意味の数学モデルによる意味的連想検索方式[3]へ適用するステップ(Step-I～Step-III)を示す．

Step-I *MDS*の生成

意味的検索空間*MDS*生成のための特徴行列として，上位下位関係行列*M*を設定する．*M*の相関行列$M^{T}M$を固有値分解し,固有ベクトルから構成される正規直交空間*MDS*を生成する．*MDS*は，ある単語についての上位概念や下位概念となる単語群を得るための状況を規定する文脈を与えることにより，その文脈に応じた単語間の方向性を有する関係の計量を可能とする意味的検索空間である．

Step-II 検索語，検索対象データベクトルの生成

上位下位関係行列の組(VS−1，VS−2，VS−3)(図3) を，検索語ベクトル，および検索対象データベクトルを形成するための特徴行列として設定する．検索語ベクトル，および

検索対象データベクトルを新規に追加する場合は，それぞれの特徴行列に追加する．

Step-III 検索語，検索対象データベクトルの*MDS*への写像

Step-II により得られた検索語(キーワードおよび文脈語）ベクトル$\mathbf{q}$，および検索対象データベクトル$\mathbf{p}$を，*MDS* 上へ写像する．検索語ベクトル$\mathbf{q}$の*MDS*への写像ベクトル$\mathbf{q}0$を，無限大ノルムで正規化することにより，意味重心ベクトル$\mathbf{G}+$を得る．$\mathbf{G}+$は，意味的検索空間*MDS*における部分空間選択に用いられる．$\mathbf{G}+$により選択された部分空間上に$\mathbf{p}$を写像し，そのノルムを計量することにより，検索語列と検索対象データ$\mathbf{P}$の相関量計算が行われる．

3.4 問い合わせ処理の計算量

問い合わせを行う前提として，意味的検索空間 *MDS* の生成を行う．*MDS* の生成は，行列の固有値計算を行うため，ベクトル次元数に応じた大きな計算量を必要とするが，上位下位関係行列を定義した後の *MDS* の生成は初期的に一回だけ行えばよい. *MDS* への任意の検索対象データをマッピングすることにより，意味的連想検索を適用することができる．このとき，*MDS* の生成は必要なく，また，*MDS* の生成時間に比べて大変短い．問い合わせ処理は，大きく部分空間選択処理と，検索対象データのノルム計算処理からなる．部分空間選択処理の計算量は, *MDS* へマッピングされた検索対象データ数に依存しておらず，ノルム計算時間と比較して大変短い．ノルム計算の計算量は，検索対象データ数 N のオーダ，すなわち $O(N)$ であり，ノルム計算結果をランキングする場合は，そのランキング(ソーティング)にかかる計算量 $O(N \log N)$ が必要になる．このノルム計算について，提案方式では，文献[4]に示す高速化アルゴリズムにより，$O(N)$ 以下の計算によりノルムの大きい(高い相関を有する)検索対象データを獲得することが可能となる．

4. 実験

4.1 実験環境

本実験では，3 章で示した提案方式により，宇宙開発事業団(旧 NASDA, 現 JAXA)が所有する安全信頼性情報 95 件を対象として，「上位概念検索」，および「下位概念検索」可能である意味的連想検索システム(表 1)を実現した．意味の数学モデル[3]へ適用するための特徴行列作成のために, 宇宙開発事業団で提供している文書に掲載されている安全信頼性情報に関する事象単語群(表 2)を用いた．

表 1 意味的連想検索空間の構成
Table 1 Description of the system

Feature 数	基本データ数	空間次元数
366	366	363

表 2 システム実現に用いた単語群の例
Table 2 Examples of terms for system realization

大分類項目 (上位概念)	中分類項目 (中位概念)	小分類項目 (下位概念)
機械的性能	質量特性不良	重心規格外、偏心 質量規格外 ・・・
	回転異常	停止位置・角度規格外 回転数規格外 ・・・
光学的性能	視野視線異常	視野範囲異常 ・・・

この分類表に示されている単語群は，(a) 開発・製造から射場整備時までのロケット及び人工衛星の不具合についての単語が示されており，(b) 大分類，中分類，小分類の 3 階層のツリー構造を持っている．検索対象文書のメタデータ設定（表 3〜5）は，宇宙工学の専門家が行った．

4.2 実験目的・方法

本実験では，実現した意味的連想検索システムが，検索語の「上位概念」，「下位概念」語に関する文書について，それぞれ独立な検索を実現していることを確認する．以下，それぞれの文書検索を「上位概念検索」，および「下位概念検索」とする．上位概念検索においては，検索語の「上位概念」について記述のある内容の文書について相関量が高く算出され, 検索結果の上位にランクされることを確認する．また，下位概念検索においては，検索語の「下位概念」について記述のある内容の文書について相関量が高く算出され，検索結果の上位にランクされることを確認する．

本実験では，問い合わせとして，中位概念の事象単語である「絶縁不良」を用いて，それぞれ上位概念，下位概念の検索を行う．また，最下位概念の事象単語である「接触不良」を用いた上位概念検索，および最上位概念の事象単語である「材料・部品不良」を用いた下位概念検索を行う．

4.3 実験結果・考察

実験結果を表 3〜5 に示す．表中の文書のメタデータでは，{数字}は，検索語について概念関係の階層を表している．例えば，{1}は検索語の上位概念の単語，また{3}は，下位概念の単語であることを表している．また太字の事象単語は，検索語と同一の上位概念を持つ単語であることを表している．

表 3 に示す結果では, 中位概念の事象単語である問い合わせ「絶縁不良」を用いて，それぞれ上位概念検索，下位概念検索をした上位 10 件の結果，および文書のメタデータ(上位 2 件)を示している．上位概念検索については､メタデータとして,「絶縁不良」の上位概念である「電気的性能」が設定されている文書 CR-58204(2 位)が上位に検索されている. また，下位概念検索については，メタデータとして,「絶縁不良」の下位概念である「短絡（ショート)」や「絶縁低下」が設定されている文書 CRA-99004(1 位)および CRA-99006(2 位)が上位に検索されている. この実験結果は，提案方式では，検索語の「上位概念」，「下位概念」に関する，それぞれ独立な相関量計算を行い，文書検索を実現していることを示している．

表 4 に示す結果では, 最下位概念の事象単語である問い合わせ「接触不良」を用いて，上位概念検索をした結果を示しており,「接触不良」の上位概念である「絶縁不良」がメタデータとして設定されている文書 CRA-97002A(1 位)とともに，最上位概念である「電気的性能」がメタデータとして設定されている文書 CR-58204(2 位)および CR-59003X(6 位)が上位に検索されている．この実験結果は，提案方式において「上位概念」検索を行った場合に，1 つ上の階層のみではなく，より上位の階層の上位概念について文書検索を実現していることを示している．表 5 に示す結果では，最上位概念の事象単語である問い合わせ「材料・部品不良」を用いて，下位概念検索をした結果を示しており，同様に「下位概念」検索を行った場合に，1 つ下の階層のみではなく，より下位の階層の下位概念について文書検索を実現していることが確認できる．

以上の実験結果により, 提案方式により実現した意味的連想検索システムが，「上位概念」，「下位概念」に関する，それぞれ独立な文書検索を実現していることが確認できる．なお, 本実験では, 3 階層の上位下位概念ツリーを用いたが,

提案方式は多階層のツリーにも適用可能である．この場合，より多階層の上位下位概念に関する文書検索を行う場合は，3 章の上位関係行列 (*Mp*)，下位関係行列(*Mc*)を用いて，元の問い合わせベクトルに対して，その上位概念語のベクトルまたは下位概念語のベクトル再帰的に合成することにより，問い合わせ拡張を行うことにより実現できる．

表 3「絶縁不良{2}」の検索結果
Table 3 Results of “Insulation trouble”

順位	上位概念検索		下位概念検索	
	文書 ID	相関量	文書 ID	相関量
1	CR-58102	0.3607	CRA-99004	0.4002
2	CR-58204	0.3574	CRA-99006	0.3111
3	CR-59003X	0.3455	CR-58113	0.2898
4	RIS-61-003	0.3123	CR-59204	0.2815
5	CR-58603	0.3001	CR-58102	0.2750
6	CR-58201	0.2963	CR-58905	0.2612
7	CR-58111	0.2931	CRA-98001A	0.2561
8	CR-59107	0.2652	CR-59106	0.2516
9	CR-58905	0.2638	CR-58210A	0.2479
10	CR-59401	0.2615	RIS-61-003	0.2310

文書のメタデータ（上位概念検索：上位 2 件）

文書 ID	メタデータ
CR-58102	「絶縁不良{2}」 「短絡（ショート）」 「技量不足」 「実装設計不十分」・・・
CR-58204	「電気的性能{1}」 「動作不安定」 「導通不良」 「抵抗値不良（大小）」・・・

文書のメタデータ（下位概念検索：上位 2 件）

文書 ID	メタデータ
CRA-99004	「短絡（ショート）{3}」「絶縁低下{3}」 「動作せず、始動不能、停止不能」・・・
CRA-99006	「絶縁不良{2}」 「絶縁低下{3}」 「短絡（ショート）{3}」・・・

表 4「接触不良{3}」の上位概念検索結果
Table 4 Results of “Contiguity defect”

順位（相関量）	文書 ID	メタデータ
1 位 (0.4571)	CRA-97002 A	「導通不良{2}」「接触不良」 「動作せず」・・・
2 位 (0.4321)	CR-58204	「導通不良{2}」「電気的性能{1}」・・・
6 位 (0.3744)	CR-59003X	「電気的性能{1}」・・・

表 5「材料・部品不良{1}」の下位概念検索結果
Table 5 Results of “Material and parts defect”

順位（相関量）	文書 ID	メタデータ
1 位 (0.2256)	CR-58106	「スクリーニング不完全(国産品){3}」 「スクリーニング不完全(輸入品){3}」・・・
2 位 (0.1968)	CR-59003	「国産品{2}」・・・
3 位 (0.1967)	CRA-99002	「輸入品{2}」「輸入品その他{3}」・・・

5. まとめと今後の課題

本稿では，方向性を有する関係を計量する意味的連想検索方式を示した．宇宙工学分野の単語群について上位下位概念関係を計量する意味的連想検索空間を構築し，実験により「上位概念」，「下位概念」それぞれ独立な文書検索を実現可能なことを確認した． 提案方式は，方向性を有する関係がある各種データ群を対象とした相関量計量する場合に有効である．単語間のオントロジーにおける上位下位概念関係など，広く用いられているツリー構造をともなった知識表現をベクトル表現することにより，上位下位概念、因果関係のような単語間における方向性を有する意味的関係の計量を行なうベクトル空間を生成し，上位下位概念の関係および因果関係の計量を組み合わせた様々な検索目的に応じた意味的連想検索を実現することが可能になる．

今後の課題として，このような方向性を有する関係の計量が重要な応用分野を設定し，それらの応用における空間構築方法、関連性の計量方法を実現していく予定である．

[謝辞]

本研究に関して，多くの貴重なご助言を頂いた慶應義塾大学図子泰三氏，宇宙航空研究開発機構但田育直氏，波内みさ氏，および(株)翔エンジニアリング仲春男氏にこの場を借りて感謝申し上げます．

[文献]

[1] Baeza-Yates, R. and Ribeiro-Neto, B.:``Modern Information Retrieval, " ACM Press, (1999).

[2] Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R. A.: “Indexing by latent semantic analysis”, Journal of the Society for Information Science, vol.41, no.6, pp.391-407 (1990).

[3] Kiyoki, Y., Kitagawa, T. and Hayama, T.:``A metadatabase system for semantic image search by a mathematical model of meaning," ACM SIGMOD Record, Vol.23, No. 4, pp.34-41, (1994).

[4] 宮川明子,清木 康,宮原隆行,北川高嗣: ``画像データベースを対象とした意味的連想検索の高速化アルゴリズム," 情報処理学会論文誌：データベース，Vol.41，No. SIG 1(TOD5), pp.1-10, (2000).

[5] 鷹野孝典，図子泰三，清木康: ``事象間の因果関係を扱う動的な文脈解釈機能を有する意味的連想検索方式の実現", 情報処理学会論文誌：データベース，Vol.46，No. SIG 5(TOD25), pp.40-55, (2005).

[6] Salton. G.: ``The SMART Retrieval System -- Experiments in Automatic Document Processing," Prentice Hall Inc., Englewood Cliffs, NJ, (1971).

[7] Salton. G.: ``Automatic Indexing and Abstracting," Document Retrieval Systems, Taylor Graham Series In Foundations Of Information Science, pp.42-80, (1988).

[8] 図子泰三，清木 康，鷹野孝典，但田育直，波内みさ: “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”，情報処理学会論文誌：データベース，Vol. 45, No. SIG 7(TOD22), pp.124-136, (2004).

鷹野 孝典 Kosuke TAKANO

慶應義塾大学政策・メディア研究科博士課程在学中．2003 慶應義塾大学大学院政策・メディア研究科修士課程修了．データベースシステムの研究に従事．情報処理学会学生会員．日本データベース学会学生会員．

清木 康 Yasushi KIYOKI

慶應義塾大学環境情報学部教授．1983 慶應義塾大学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．1984～1996 筑波大学電子・情報工学系講師，助教授を経て，1996 慶應義塾大学環境情報学部助教授，1998 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，情報処理学会各会員．