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

An implementation method of semantic associative search spaces for medical documents

河本 稔[†]　清木 康[††]　吉田 尚史[†]
藤島 清太郎[‡]　相磯 貞和[‡]

Minoru KAWAMOTO　Yasushi KIYOKI
Naofumi YOSHIDA　Seitaro FUJISHIMA
Sadakazu AISO

本稿では，医療分野に関する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を示す．本方式は，特定の専門分野を対象として，主に単語の直接的な説明が記述されている事典を用いて，その分野の情報源を対象とした意味的連想検索を実現するための検索空間を生成する方式である．本方式は，既に提案されている意味的連想検索の医療への応用において，単語の意味の定義が明示的でなく，主に，その単語の説明を記述している事典を対象とした意味空間の実現を可能とする．

In this paper, we present an implementation method of a semantic associative search space for medical documents. This method is used to create a semantic associative search space for a domain-specific field. This method realizes space creation by using encyclopedias which describe not only word definitions but also word explanations. Our method makes it possible to create a medical semantic space by using a medical encyclopedia which includes explicit word explanations.

1. はじめに

現在，医療に関するさまざまな電子ドキュメント群が存在している．これらのドキュメント群はそれぞれ検索者の視点に応じて異なった情報を持っていると考えられる．例えば，検索者が医師であるか，患者であるかといった検索者の違いに応じた視点があり，検索者の関心や目的に応じた視点が存在する．一般にドキュメント群は，検索者とその利用目的によって異なる多様な意味を持っている．視点によって異なる意味を持つと考えられるドキュメント群を対象とし，検索者の文脈に沿ったドキュメント検索の手法として意味の数学モデル[1],[2]によるドキュメント検索方式[3],[4]が提案されている．

被説明語	説明語群
lung	either of two breathing organs in the chest of humans or certain other animals
cough	to push air out from the throat suddenly , with a short rough sound , [esp.] because of discomfort in the throat during a cold or other infection

図1 Longman Dictionary of Contemporary English[5]における単語の説明の例

被説明語	説明文
咳・痰	「咳と痰はほとんどの呼吸器疾患でみられる．(中略) 咳は，気道に侵入した異物を除去するための重要な生体防御機構であり，(中略) 咳の発生機序は，副交感神経の分布しているところを刺激すれば」(略)

図2 特徴語のみによらず，また，定義のみに限定されない記述による単語の説明の例(『今日の診療[7]』より)

検索対象とする分野の要素となる特徴語群(feature words)によって単語の意味の定義が示されている辞典から一般的な分野を対象とした意味的連想検索空間(以下，意味空間)を自動的に構成する方式[1]がすでに提案されている．

それに対して本方式は，特定の専門分野を対象として，主に単語の説明を記述した事典により意味空間を構成する方式である．

意味空間の実現について，Longman Dictionary of Contemporary English[5] のように単語の意味の定義を検索対象の分野の要素となる特徴語群により説明している辞典が利用可能な場合(図1)においては，その辞書を利用して自動的に意味空間を構成することが可能である[2]．そのような辞書が利用可能でない場合(例として図2を参照)においては，参考となる文献を参照し多数の単語について意味を記述するという多くのワークロードをこなす必要性があり，客観的でかつ性能の高い意味空間を構成することが困難となる．

意味空間を自動的に生成可能な辞典の条件は，Longman Dictionary of Contemporary Englishに代表されるように，(1)単語の説明として単語の定義について記述が行われていること(2)特徴語群のみによる記述が行われていること．の2点である．本方式は，このような条件を満たしていない事典，すなわち，(1)単語の説明として単語の定義のみではなく関連する事例などについて広範な説明を持ち，定義の説明が十分でないもの，(2)単語の説明に利用する特徴語群を選別せずに記述が行われているものを参照した意味空間の実現を対象とする．

本方式で対象としている意味的連想検索方式[1],[3]は，多変量解析による空間生成を用いた情報検索方式（例えばLSI[6]）とは次の点で本質的に異なる．本方式が対象とする意味的連想検索方式[1],[3]では，直交空間における部分空間の選択を行う演算を定義し，その演算により，言葉と言葉，あるいは，言葉とメディアデータ（例えばドキュメントデータ）の間の意味的な関係を，与えられた文脈に応じて動的に計算することが可能[1],[3]となる．

本方式は，単語の定義のみではなく，関連する事例などについて広範な説明を持ち，かつ，単語の説明に利用する特徴語群を選別せずに記述が行われている事典を参照した意味

† 学生会員　慶應義塾大学政策・メディア研究科修士課程
minoru@mdbl.sfc.keio.ac.jp
†† 正会員　慶應義塾大学環境情報学部
kiyoki@sfc.keio.ac.jp
† 正会員　慶應義塾大学政策・メディア研究科
naofumi@sfc.keio.ac.jp
‡ 慶應義塾大学医学部　fujishim@sc.itc.keio.ac.jp
‡ 慶應義塾大学医学部　aiso@sc.itc.keio.ac.jp

基本データ($w_1 \dots w_m$)	特徴語列($f_1 \cdots f_n$)
気道狭窄	乾性ラ音 吸気性
気道異物	乾性咳 気道狭窄 呼気延長 呼吸困難 喘鳴
気道閉塞	呼気延長
閉塞性換気障害	呼気延長
肺線維症	から咳 ベルクロ・ラ音 横隔膜挙上 横隔膜直上蜂窩肺 拡散障害 拘束性換気障害 拘束性肺疾患 浅く速い呼吸 線状影 捻髪音 肺拡散能低下 閉塞型無呼吸症候群 網状影 粒状影

図3 本方式により生成された基本データセットの例(抜粋)

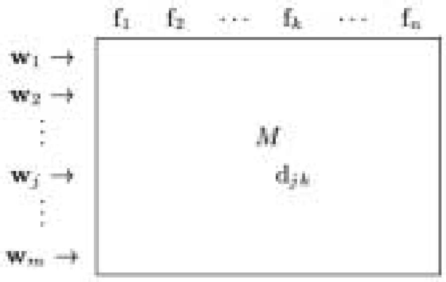

図 4 基本データ行列

空間を構成する方式である．本方式により，専門分野におけるドキュメント群を対象とした専門性の高い検索環境を検索者に提供することが可能となる．

本稿では，医療分野における適用を例として，本方式による意味空間の実現について示す．さらに，本稿では，実験により本方式の実現可能性と有効性を示す．

2. 意味的連想検索空間の実現方式

専門的な知識が文献により提供されている医療の分野において意味空間を構成するために，事典からの情報抽出を行うことによって意味空間を構成する．

意味空間を構成する際に，もっとも基本的な構成要素となるのは特徴語群である．意味空間を構成する単語群は，すべてこの特徴語群から特徴づけを受けることによって意味空間における意味付けが決定され，意味空間上に配置される．検索語と検索対象メタデータの相関の度合いは，この意味空間上におけるベクトル同士の演算として実現される．

特徴語群は，より多くの単語が意味空間において利用可能となる環境を提供するために，できる限り多くの，理想的にはすべての単語を効率的に説明できる単語の集合であることが求められる．

意味空間を構成するには，特徴語のほかに基本データと呼ぶ単語セットを必要とする．基本データは，特徴語による説明を行い，特徴語と基本データとの基本データセットを形成する．基本データセットは図3に示すように構成される．
本方式で扱っている基本データセットとは，基本データに対して特徴語群から，基本データと関連のある単語群を基本データ単語の意味の定義として表現したものである．基本データ群を列挙し，基本データと関連がある特徴語を基本データの説明語列として記述することで基本データ群を特徴語群を用いて特徴づけするものである．

2.1 実現方式

意味的連想検索において利用する基本データセットの生成を，次の手順により行う．

Step-1　利用する事典の情報を参照し，主に見出しなどから基本データの候補となる単語群を抽出する．候補となる単語群は，事典において見出しとなっていて症状や治療方法などについて述べられている単語である．基本データとなる単語群は，特徴語による説明を行うことによって意味空間における特徴語の意味付けを決定する働きがある．基本データの数が不足した場合，意味空間において特徴語が表すことのできる意味が減少してしまう結果となる．そのため，この Step においては網羅的に多くの基本データを抽出する．

Step-2 ***Step-1*** において選んだ基本データの候補群についてのその単語の意味を説明している文を事典から選ぶ．選ばれた記述文に対して形態素解析を行い，機能語を取り除き，特徴語の候補とする．本方式は単語の意味の定義を表現することが目的であるため，この Step においては単語の意味の定義と関係のない記述を適切に排除することが重要となる．

Step-3 抽出された単語群から，(1)**説明頻度**，および(2)その分野において判断される**重要度**を考慮し，特徴語として必要であると考えられる単語を選択する．

(1)**説明頻度**．２種類以上の単語の説明としての出現を基本的な基準とする．一般に頻度の高いものは特徴語としての重要度の高い傾向がある．

(2)**重要度**．特徴語を選ぶ際の重要度の基準はより多くの概念を特徴付けるために，必要と思われる普遍性の高い単語であることを選択の基準とする．

Step-4 ***Step-1*** において選んだ基本データ候補群の単語を，***Step-3*** において選んだ特徴語群によって説明するベクトルを作る．事典の記述内容を参照し，基本データ群を特徴語群により特徴づけ，基本データセットを生成する．

2.2 医療分野のドキュメント検索

本方式は，医療分野のドキュメント群を対象とした，文脈に応じた動的な相関量計算機能を適用する意味空間を実現する．医療分野のドキュメント群を対象とした意味的連想検索により，特定の病名や症状と関連のある文書を検索するためのシステムが提供可能となる．

また，本方式は医療分野における診断を支援するシステムへ応用可能であると考えられる．本方式は医師による診断の信頼性を向上させるシステムの基礎となる方式である．

3．提案方式の実現

文献[1],[3]において提案されている意味の数学モデルで用いられる基本データ群，特徴語群の基本データ行列を本方式により構成する方法について述べる．また，得られた基本データ群と特徴語群の基本データ行列を文献[1],[3]において提案されている意味の数学モデルによる意味的連想検索方式へ適用する方法を示す．

3.1 意味の数学モデルへの適用

基本データ群と特徴語群の基本データセットを対象として文献[1],[3]においてすでに提案されている意味の数学モデルによる意味的連想検索方式において用いられるメタデータ空間を生成する方法を示す．

⑴基本データセットからのメタデータ空間の生成

基本データセットから基本データ行列を設定する．基本データ行列において，基本データ単語は行(w_j)に，特徴語は列(f_k) に相当する(図 4)．基本データ行列中の対応する特徴語の d_{jk} の値は，基本データ単語 w_j のもつ特徴語群中に該当する列 k の特徴語が存在する場合は‘１’を，存在しない場合

	相関度	ドキュメント名	ドキュメントのメタデータ
1	0.693848	特発性肺線維症	肺線維症 間質性肺炎
2	0.575438	膠原病肺	膠原病肺 間質性肺炎
3	0.492448	急性間質性肺炎	急性間質性肺炎 間質性肺炎 呼吸困難
4	0.343448	在宅酸素療法	慢性閉塞性肺疾患 COPD 肺結核後遺症 間質性肺炎 肺線維症
5	0.332587	慢性呼吸不全の日常管理	慢性閉塞性肺疾患 COPD 肺結核後遺症 間質性肺炎
6	0.055973	肺血栓, 肺塞栓	肺血栓塞栓症 呼吸困難 肺梗塞
7	0.052506	過換気症候群	過換気症候群 呼吸困難
8	0.052002	胸痛	気胸 肺梗塞 胸膜炎
9	0.048357	急性呼吸促迫症候群	ARDS 誤嚥 ウイルス肺炎 カリニ肺炎 細菌性肺炎
10	0.034042	肺真菌症	肺アスペルギローマ

実験 A の結果(文脈語：間質性肺炎)

	相関度	ドキュメント名	ドキュメントのメタデータ
1	0.551320	喘息:喘息発作への対応	呼吸困難
2	0.425764	喘息:慢性喘息の管理	気道閉塞 呼吸困難 咳
3	0.420628	過換気症候群	過換気症候群 呼吸困難
4	0.419918	過敏性肺炎	過敏性肺炎
5	0.416797	特発性器質化肺炎	器質化肺炎 BOOP 呼吸困難 浸潤影
6	0.409892	肺好酸球性肉芽腫症	呼吸困難 咳
7	0.378254	急性間質性肺炎	急性間質性肺炎 間質性肺炎 呼吸困難
8	0.375352	肺血栓，肺塞栓	肺血栓塞栓症 呼吸困難 肺梗塞
9	0.311113	薬剤誘起性肺炎	肺炎 呼吸困難 咳
10	0.308584	肺炎（院内感染）	肺炎 誤嚥 咳 呼吸困難 浸潤影

実験 B の結果(文脈語：気道異物)

	相関度	ドキュメント名	ドキュメントのメタデータ
1	0.367435	慢性呼吸不全の日常管理	慢性閉塞性肺疾患 COPD 肺結核後遺症 間質性肺炎
2	0.363439	在宅酸素療法	慢性閉塞性肺疾患 COPD 肺結核後遺症 間質性肺炎 肺線維症
3	0.335933	呼吸器疾患の動向	慢性閉塞性肺疾患 COPD
4	0.318378	慢性閉塞性肺疾患	慢性閉塞性肺疾患 肺気腫 慢性気管支炎 COPD
5	0.191669	びまん性汎細気管支炎／びまん性気管支拡張症	気管支拡張症 閉塞性換気障害
6	0.179777	肺結核	肺結核 結核
7	0.176727	CO 2 ナルコーシス	CO2 ナルコーシス 呼吸困難 チアノーゼ 神経筋疾患 気胸
8	0.111122	嚥下性肺炎	肺炎 誤嚥
9	0.108581	肺癌	肺癌
10	0.106143	肺炎（院外感染）	マイコプラズマやクラミジアによる肺炎 肺炎 細菌性肺炎

実験 C の結果(文脈語：珪肺)

は‘ 0 ’を，否定的な意味で存在する場合は‘ -1 ’を付与する．全基本データ単語について対応する行ベクトルを生成し，基本データ行列とする．

⑵ドキュメントデータのメタデータをメタデータ空間へ写像

メタデータ空間へドキュメント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．ドキュメントデータ D には，メタデータとして t 個の基本データ $w_1, w_2, \ldots, w_t$ が以下のように付与されていることを前提としている．

$$D = \{w_1, w_2, \ldots w_t\}. \quad (1)$$

各基本データは，ベクトル表現された特徴を持っている．

$$W_i = (f_{i1}, f_{i2}, \ldots f_{in}). \quad (2)$$

各ドキュメントデータは，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，メタデータ空間へ写像される．

⑶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，ドキュメント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，各ドキュメント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

ここでは本方式の有効性を検証するために行った実験の結果を示す．『今日の診療 Vol.12』[7]を参照し，肺・呼吸器の分野を例として対象とした意味空間を構成した結果，基本データ群 217 語，特徴語群 219 語が抽出された．抽出した基本データ群と特徴語群から，基本データセットを生成した(図 3)．意味的連想検索方式[1][3]を適用した意味空間生成の結果として，149 次元の意味空間を生成した．

4.1 実験:検索性能に関する実験

本方式により構成された意味空間と医療に関連するドキュメント群を用いて検索を行い，検索結果を検証する．

例として文脈語“ 間質性肺炎 ”を指定し，検索結果の妥当性の評価を行った(実験A). 検索結果は図5の通りとなった．同様に，文脈語“ 気道異物 ”を指定し，検索を実行した結果(実験 B)は図 6 に，“ 珪肺 ”を指定し，検索を実行した結果(実験 C)は図 7 に表した出力を得た．

4.2 実験に関する考察

実験 A の検索結果においては，文脈語の“ 間質性肺炎 ”自体をメタデータとして持つドキュメントが上位に検索されている．文脈語と検索対象のメタデータが一致している場合に，文脈語と，ドキュメントのメタデータが関連している特徴語が共通しているという理由により，上位に検索されることとなる．実験 B の検索結果において，文脈語と一致しているメタデータを持っているドキュメントは存在しないが，文脈語の“ 気道異物 ”と，特徴において関連のある“ 呼吸困難 ”をメタデータとして有するドキュメントが上位に検索され

ている．これらの単語に共通する特徴語の“ 呼吸困難 ”は，**Step-3** において(1)の出現頻度を基準に選定されたものである．実験 C の検索結果において，文脈語と一致しているメタデータを持っているドキュメントは存在しないが，文脈語の“ 珪肺 ”と，特徴において関連のある“ 肺結核後遺症 ”をメタデータとして有するドキュメントが上位に検索されている．これらの単語に共通する特徴語の“ 肺結核 ”は，**Step-3** の特徴語の選定において(2)の重要度を基準に選定(説明頻度:1)されたものである．

実験 A において, 文脈語と共通のメタデータを有するドキュメントが上位に検索された．実験 B において，**Step-3**(1)で示された説明頻度の基準により選出された特徴語の影響により，特徴において関連のある“ 呼吸困難 ”をメタデータとして有するドキュメントが上位に検索されている．実験 C において，**Step-3**(2)で示された重要度の基準により, “ 珪肺 ”と“ 肺結核後遺症 ”が共通に持つ“ 肺結核 ”という特徴語により関連づけられ ，“ 肺結核後遺症 ”をメタデータとして持つドキュメントが, “ 珪肺 ”という文脈語として与えることにより獲得されている．

以上の実験により，妥当性の高い，意味的に正しい単語の特徴づけを特徴語群により設定可能であることを示した．本実験により，生成した意味空間が意味的連想検索のための空間として機能していることが判る．

今後，より高い性能の意味空間を構築するためには，特徴語の選択や基本データの特徴づけを行う際により専門性の高い検討を加える必要があると考えられる．

5．まとめ

医療分野のドキュメント群を検索可能な環境を提供する意味的連想検索方式の意味空間を，事典から生成する方法について述べた．また，意味的連想検索方式を医療分野のドキュメント群に応用することを可能としたことで，医療ドキュメントを扱う応用的なアプリケーションに適用可能なプラットフォームを提供することを可能とした．

本研究においては医療分野のドキュメントを検索対象のデータとして想定して医療分野の意味空間の構成を行ったが，本方式は対象とする分野に依存するものではなく，事典からの意味空間の生成が，様々な分野に適用できると考える．今後の課題として，実際の医療分野での実現可能性の検証，意味的連想検索空間統合方式[8]の適用による大規模検索空間による検索の実現が挙げられる．

[謝辞]

本研究に関して貴重な御助言を頂いた筑波大学電子・情報工学系北川高嗣教授に感謝申し上げます．

[文献]

[1] Kitagawa, T. and Kiyoki, Y.: “The mathematical model 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, April (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, Amit Sheth and Wolfgang 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, (1996).

[5] “Longman Dictionary of Contemporary English”, Longman, (1987).

[6] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R: “Indexing by Latent Semantic Analysis”, Journal of the American Society of Information Science, 41(6), 391-407, (1990).

[7] “今日の診療 Vol.12 ハイブリッド CD-ROM 版”，医学書院, (2002).

[8] 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), (2002).

河本 稔 Minoru KAWAMOTO

慶應義塾大学大学院政策・メディア研究科修士課程在学中．2002 慶應義塾大学総合政策学部卒業．データベースシステムの研究に従事．日本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．

吉田 尚史 Naofumi YOSHIDA

1972 年生．1996 年筑波大学第三学群情報学類卒業．1998 年同大学院修士課程理工学研究科修了．2001 年同大学院博士課程工学研究科修了．博士（工学）．2001 年より慶應義塾大学大学院政策・メディア研究科専任講師．データベースシステム，マルチメディアシステム，医学情報システムに関する研究に従事．ACM，情報処理学会各会員．

藤島 清太郎 Seitaro FUJISHIMA

1982 年慶應義塾大学医学部卒業，同大学内科研修．1988 年～1991 年米国 Stanford 大学留学．1992 年博士（医学）取得．1997 年より同大救急部講師．同大学病院で患者診療に従事する傍ら，炎症性肺疾患などの研究に従事．日本内科学会，日本救急医学会，日本呼吸器学会などの各会員．

相磯 貞和 Sadakazu AISO

慶應義塾大学医学部教授，1976 年慶應義塾大学医学部卒業，1980 年慶應義塾大学大学院医学研究科博士課程修了，医学博士，慶應義塾大学医学部助手，専任講師，Stanford 大学医学部微生物学免疫学 Post-doctoral Fellow を経て，1993 年より現職，形態形成学の研究に従事，日本解剖学会，日本内科学会，日本消化器病学会，各会員．