国際関係分野ドキュメント群を対象とした意味的連想検索のための空間生成方式

佐々木 史織† 清木 康‡ 薬師寺 泰蔵††

†慶應義塾大学法学研究科 〒108-8345 東京都港区三田 2-15-45
‡慶應義塾大学環境情報学部 〒252-8520 神奈川県藤沢市遠藤 5322

††慶應義塾大学法学部 〒108-8345 東京都港区三田 2-15-45
E-mail: †s-sas@jcom.home.ne.jp, ‡kiyoki@sfc.keio.ac.jp , ††yakushi@iips.org

あらまし 本稿では，国際関係分野のドキュメント群を対象とした意味的連想検索空間と一般辞書を用いた意味的連想検索空間とを合成し，新しい統合空間を生成する方式を示す．本方式では，第１ステップとして，国際関係分野の専門用語集を用いて，用語間の意味的関連性を計量するベクトル空間を生成する．第２ステップとして，一般辞書を用いて構築された一般語間の意味的関連性を計量するベクトル空間のマトリクスを合成する．合成において，特定分野の専門語を一般語で定義づけるプロセスと，一般語を専門語によって特徴づけるプロセスを設定する．本方式によって生成した空間を用いることにより，一般語を用いて，意味的に関連する国際関係分野の専門用語を含むドキュメントを検索すること，および，国際関係分野の用語を用いて，意味的に関連する一般語を含むドキュメントを検索することが可能となる．

キーワード 意味的連想検索，国際関係，ドキュメントデータベース

A New Creation Method of a Semantic Retrieval Space for Documents of International Relations

Shiori SASAKI† Yasushi KIYOKI‡ and Taizo YAKUSHIJI††

†Graduate School of Law, Keio University 2-15-45 Mita, Minato-ku, Tokyo, 108-8345 Japan
‡Faulty of Environmental Information, Keio University 5322 Endo, Fujisawa-shi, Kanagawa, 252- 8520 Japan
††Faculty of Law, Keio University 2-15-45 Mita, Minato-ku, Tokyo, 108-8345 Japan

E-mail: †s-sas@jcom.home.ne.jp, ‡kiyoki@sfc.keio.ac.jp, ††yakushi@iips.org

Abstract. In this paper, we present a new creation method of a semantic retrieval space for the field of International Relations (IR). This method creates an integrated metadata space for computing semantic relationships between words in an IR lexicon and in a general dictionary. The created semantic space is applied to the mathematical model of meaning which has already been proposed. This model makes it possible to dynamically compute semantic relationships between words according to a given context. With the semantic space created by this method, we can search the documents consisting of IR terms by using general words, and also those consisting of general words by IR terms.

Keyword semantic associative search, international relations, document database

1. はじめに

現在，広域ネットワーク上には膨大なドキュメントデータが存在する．その中には，政府や国際機関の公式発表，プレス・ブリーフィング，政策ステートメント，政府高官の談話，議会議事録，NGO の活動記録等，国際関係・国際政治に関連するドキュメントも多い．国際関係・国際政治の研究者にとって，これらのドキュメントデータからの的確かつ迅速な情報獲得が課題となっている．

しかし，WWW 上のサーチエンジンのカテゴリ検索をはじめ，一般的な検索エンジンは単純なパターンマッチング方式を採用しており，単語やデータ間の相関量計算といった意味の解釈を伴った検索は困難である．また，国際関係論や国際政治学の分野においてドキュメント分析の方法として従来より行われてきた内容分析 [1)~4)]または認知構造図 [5)~7)]の手法は，主にドキュメ

ントの静的な性質を対象とした知識発見の方法であり，分析者の視点や関心に応じて意味内容を多角的・動的に分析することは困難であった．

内容分析(Content Analysis)および認知構造図(Cognitive Map)は共に，60年代～70年代にかけて国際関係・国際政治の分野に導入され，公表された文書の事後的分析を通じて政策決定者の認知，態度，あるいは対外イメージなどの分析に応用されてきた．特に，世論や政党の分析には新聞や機関紙が，意見調査を行うことが困難な各国首脳の認知分析には演説や声明，交換文書，書簡などが対象ドキュメントとして用いられている．

内容分析は，ドキュメント群中の単語またはコード化された文の出現頻度を計測し，各単語・コードと各ドキュメントとの相関度を計算し，多変量解析を行うものである．一方，認知構造図は，あるドキュメントの論理構造を概念(concept)のネットワークとみなし，各コンセプト間の因果関係を+,－,0で表すことでドキュメントの著者または発言者（政策決定者）の論理経路を分析する手法である．しかし，内容分析はドキュメント群に含まれる単語間の意味的関連性について計量できず，また，認知構造図はドキュメント間の意味的な関連性について計量することが出来ない．

内容分析においては，文や文の一部をコード化するプロセスを伴うのが一般的であるが，コード化には特定問題領域に関する極めて専門的な知識が必要となる上，信頼性を確保するために複数の人間によるチェックが必要となる．そこで，大量のデータを半自動的に扱うために，機械的なパターンマッチングによる単語の出現頻度を計測する方法もとられている．しかし単語出現頻度においては，その単語がいかなる文脈において使用されているのかが反映されないという点が問題となる．たとえば，「engagement」という単語がドキュメント内で経済的な「債務・約束」という意味で使用しているのか，あるいは軍事的な「関与」を意味しているのかは計測結果に表れない．あるいは，「development」という単語が経済的な「発展・開発」を指しているのか，軍事兵器の「開発」を表しているのかを結果から読み取ることは困難である．

一方，認知構造図は，コンセプト間の関係性を示すことは出来るものの，関係の強度を量的に表すことが出来ない．また，一ドキュメントにつき一認知図を作成しなければならず，大量のドキュメント間の比較分析には有効でない．

これらの従来の手法に対し，文献 8）～14）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は，分析者の視点や分析時の文脈に応じてダイナミックに言葉およびデータの意味解釈を実現する方式であり，その意味解釈を多次元直交ベクトル空間における相関量計算によって行うことを特徴としている．この方式を国際関係分野のドキュメント群に適用することにより，国際関係分野のドキュメント群について，コード化を経ずに可能な限り原データに近い形で大量のデータの意味的連想検索およびドキュメント間の比較分析が可能になると考えられる．本稿では，意味的連想検索方式を国際関係分野ドキュメントの検索・分析に応用する環境を実現するために，意味の数学モデルを対象とした国際関係分野意味的連想検索空間の生成方式を示す．

本方式の特徴は，主に二つある．第一に，国際関係分野のドキュメント分析に応用可能な意味的連想検索空間を生成し，対象ドキュメント群の意味的相関関係を計量する機構を実現する点にある．第二に，専門用語に関する情報源（用語集）Aを対象として生成する意味的計量空間と，一般語に関する情報源（辞書）Bを対象として生成する意味的計量空間を統合し，専門用語と一般語の両者の意味的関係の計量を行うことができる意味的連想検索空間を実現する点にある．前者は国際関係論・国際政治学における方法論的意義を持ち，後者はデータ工学的な意義を持つと考えられる．

2. 意味的連想検索方式

本方式の適用対象として，文献 8)~14)に提案されている意味の数学モデルに基づくメディアデータ意味的連想検索方式の概要を示す．

2.1. 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．

2.2. メディアデータの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，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 $w_1, w_2, \cdots, w_t$ が以下のように付与されていることを前提としている．

$$P = \{ w_1, w_2, \cdots, w_t \} \quad (1)$$

各基本データは，ベクトル表現された特徴を持っている．

$$w_i = (f_{i1}, f_{i2}, \cdots, f_{in}) \quad (2)$$

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 *MDS* へ写像される．

2.3. 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.まず，メタデータ空間 *MDS* に各単語に対応するベクトルが写像され，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．次に，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.

このプロセスにより，検索者が与えたコンテク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．与えられたコンテク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，与えられたコンテク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，与えられたコンテク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，コンテク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.

2.4. メタデータ空間 *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 *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.

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各メディアデータに対応するベクトルの大きさが単位ベクトルとなるように 2 ノルムで正規化している.

3. 本方式による空間生成方式

本方式は，国際関係（International Relations:以下 IR）分野に関するドキュメントデータ群を対象とした，専門用語を含有する意味的検索空間を生成することを目的としている．まず，専門用語集を用いて基本データ行列を生成する．次にこれを，一般辞書を用いて構築された既存のベクトル空間のマトリクスに合成する．生成および合成に際して，意味空間における「関連性(reference)」と「定義（definition）」というコンセプトを基本とし，専門用語を一般語で定義づけるプロセス

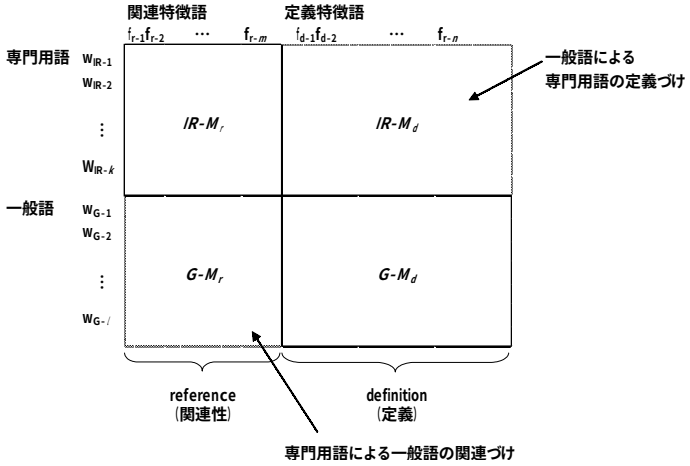

図 1 国際関係分野のメタデータ空間の構築方式

と，一般語を専門語との関連によって特徴づけるプロセスを行う.

本方式によって生成した空間を用いることにより，IR 分野に関する検索対象ドキュメントの意味的相関関係をより適切に表現することができる．すなわち，一般語を用いて IR 分野の専門用語によるドキュメントを検索すること，および，IR 分野の用語を用いて一般語によるドキュメントを検索することが可能となる.本方式は,検索対象として IR に関するドキュメント群が存在し，この分野の専門用語集と，一般的な語を説明する辞書とが存在することを前提としている.逆に，専門用語集と一般語を説明する辞書が存在するならば,他の特定分野にも応用可能な空間生成方式であるといえる.

本稿における「特徴語(feature)」とは，意味的検索空間生成のための行列において横軸にあたる単語および用語を指し,「基本語」および「基本用語」は行列において縦軸にあたる基本データを示す.IR 分野の基本用語は w_{IR} と表し,一般の基本語を w_G と表すこととする.

また特徴語のうち，基本語との「関連性」を表す特徴語は「関連特徴語(f_r)」として定義し，基本語の定義を表す特徴語を「定義特徴語(f_d)」と定義する.

図 1 は，空間生成方式の構成を示している．*IR-M*$_r$ は *k* 個の IR 基本用語(w_{IR-1}, w_{IR-2}, …, w_{IR-k})について *m* 個の IR 分野の関連特徴語(f_{r-1}, f_{r-2}, …, f_{r-m})で特徴づけた，IR 用語間の関連性を示したメタデータ行列であり，*G-M*$_d$ は *l* 個の一般基本語(w_{G-1}, w_{G-2}, …, w_{G-l})について一般の定義特徴語（f_{d-1}, f_{d-2}, …, f_{d-n}）で定義を示した

メタデータ行列である．ここに，IR 基本用語について定義特徴語で特徴づけを行う部分 $IR\text{-}M_d$，および，一般語について IR 関連特徴語で関連づけを行う部分 $G\text{-}M_r$ を加えることにより，統合メタデータ行列 $IR/G\text{-}M_{rd}$ が生成される．

3.1. 国際関係分野の基本データ行列の生成

行列 $IR\text{-}M_r$ を生成する．第一に，IR 分野を表現するために必要な特徴語群を準備する．専門用語集を用いて，各項目の説明文の中から関連する他の項目を抽出し，この集合を関連特徴語群とする．第二に，同用語集を用いて各項目を抽出し，この集合を基本用語群とする．第三に，関連特徴語群を用いて各基本用語の特徴づけを行う．同用語集を用いて，各基本用語の説明文に現れる関連特徴語には 1 を，現れない関連特徴語には 0 を，否定的な意味で現れる関連特徴語には-1 を設定する．以上のプロセスにより，IR 分野における基本用語と関連特徴語の関係を示す基本データ行列が生成される．

3.2. 一般語辞書によるデータ行列との合成

行列 $IR\text{-}M_r$ と行列 $G\text{-}M_d$ を合成するため，部分 $G\text{-}M_r$ と部分 $IR\text{-}M_d$ を生成する．

（１）専門語による一般基本語の関連づけ

部分 $G\text{-}M_r$ を生成する．すなわち，l 個の一般基本語($w_{G\text{-}1}$, $w_{G\text{-}2}$, …, $w_{G\text{-}l}$)について，m 個の IR 分野の関連特徴語($f_{r\text{-}1}$, $f_{r\text{-}2}$, …, $f_{r\text{-}m}$)で特徴づける．

（２）一般基本語による専門語の定義づけ

部分 $IR\text{-}M_d$ を生成する．すなわち，k 個の IR 基本用語($w_{IR\text{-}1}$, $w_{IR\text{-}2}$, …, $w_{IR\text{-}k}$)について，一般の定義特徴語（$f_{d\text{-}1}$, $f_{d\text{-}2}$, …, $f_{d\text{-}n}$）で特徴づける．

（３）その他の語の追加

行列 $IR\text{-}M_r$ と行列 $G\text{-}M_d$ のいずれにも存在しないが検索対象ドキュメント群に頻出する語を，基本データとして縦の列に追加し，IR の関連特徴語と一般語の定義特徴語で特徴づける．

以上の生成･合成プロセスにより，統合行列 $IR/G\text{-}M_{rd}$ が生成される．

4. 本方式の実現

4.1. 基本データ行列との合成

本方式による実現例として，汎用されている IR 用語集の ”Dictionary of International Relations ”[10]（以下，” IR-Dic. ” と呼ぶ）を用い，3.で示した方法で行列を生成した．この用語集では，716 の専門用語（項目）について，その定義，出典，歴史および他の用語（項目）との関連性を説明している．そのうち，716 の各項目の説明文から関連項目のみを関連特徴語 f_r として抽出し，3.1 で示した方法によって値を決定し，IR 基本データ行列とした．例えば，「arms control（軍備管理）」という項目については，「capability」「actor」「crisis management」「deterrence」「disarmament」「Cold War」「superpower」「non-proliferation」「ABC weapons」「security regime」などの関連特徴語に 1 という値が与えられる．この基本行列 $IR\text{-}M_r$ は，IR-Dic.内の項目間の関連性を表すものであり，基本データ数 712，特徴語数 712 の 712×712 行列となった．なお，この行列のベクトルを基に生成した意味的連想検索空間（IR 空間）は，次元数 710 となった．

4.2. 一般語データ行列との合成

生成した行列 $IR\text{-}M_r$ に，一般語辞書を用いて構築された既存の行列 $G\text{-}M_d$ を合成する．$G\text{-}M_d$ は，英英辞書である ” Longman Dictionary of Contemporary English ”[11]（以下，” Longman-Dic. ” と呼ぶ）を用いて生成されている．Longman-Dic.は約 56000 語の一般語について約 2000 語の基本単語で説明した辞書である．$G\text{-}M_d$ はこの約 2000 語の基本単語を他の基本単語で定義づけて生成された約 2000×約 2000 の行列であり，各行は一般語の定義特徴語 f_d によって特徴づけられた一般基本単語のベクトルを表している．

（１）専門語による一般語の関連づけ

3.2.(1)に示した方法で，部分 $G\text{-}M_r$ を生成する．例として，一般語の基本語「arms」は，「arms control」「arms race」「arms sales」といった IR 関連特徴語で特徴づけられる．特徴づけに際しては，IR 分野の専門家の知識によるチェックを行った．

（２）一般語による専門語の定義づけ

3.2.(2)に示した方法で，部分 $IR\text{-}M_d$ を生成する．例として，IR の基本用語「arms control」は，「arms」「control」「reduce」「remove」「weapon」「threat」「force」などの定義特徴語で特徴づけられる．特徴づけは，IR-Dic.の説明文中から用語の定義の部分について動詞と名詞を抽出した．定義特徴語にない単語がある場合は Longman-Dic.で調べ，動詞と名詞を抽出した．

(３)その他の基本語の追加

行列 $IR\text{-}M_r$ と行列 $G\text{-}M_d$ のいずれにも存在しないが検索対象ドキュメント群に頻出する語，たとえば「democracy」「economy」「policy」といった重要単語を基本データに追加し，Longman-Dic.および IR-Dic.を用いて定義特徴語と関連特徴語で特徴づけた．

以上の作業の結果，基本データ数約 2000+712，特徴語数 2861 の行列が生成された．なお，この行列を基に生成された意味的連想検索空間（統合空間）の次元数は 2846 であった．

5. 実験

本方式による統合空間の実現可能性および有効性を検証するため，次の実験を行った．

実験 1 IR 空間と統合空間における単語間相関量比較

実験 2 統合空間におけるドキュメント検索適用実験

5.1. 実験 1

5.1.1. 実験方法

実験 1-1 IR 専門用語である「arms control」を検索語として選択し, 4.1 で生成した IR 空間と 4.2 のプロセスを経て生成した統合空間における相関量の高い

表 1 実験 1-1 IR 検索語→IR 用語（空間比較）

	IR基本空間		総合空間	
順位	単語	相関量	単語	相関量
1	NPT	0.422383	**weapon**	0.597043
2	arms control	0.386465	NPT	0.411523
3	nuclear proliferation	0.367211	nuclear weapons	0.401119
4	inspection	0.354512	nuclear proliferation	0.370773
5	non-proliferation	0.345976	**INF treaty**	0.35418
6	proliferation	0.344098	non-proliferation	0.346241
7	nuclear weapons	0.32252	tactical nuclear weapons	0.332596
8	preventive war	0.315053	**START II**	0.331209
9	second strike	0.30641	CTBT	0.321286
10	non offensive defence	0.305555	Cuban missile crisis	0.312302
11	MAD	0.30432	proliferation	0.308877
12	deterrence	0.303463	**START I**	0.305525
13	CTBT	0.303237	**SALT**	0.297272
14	Cuban missile crisis	0.29737	**weapons of mass dest**	0.294429
15	accidental war	0.290768	**cruise missile**	0.294311
16	limited nuclear war	0.288547	second strike	0.285025
17	parity	0.287821	**massive retaliation**	0.27672
18	force	0.283384	arms control	0.268601
19	realism	0.279878	**missile**	0.268427
20	verification	0.277056	deterrence	0.266044
～	～	～	～	～
29	tactical nuclear weapons	0.261701	flexible response	0.242601
30	chemical and biological wa	0.260292	horizontal proliferation	0.23927

単語上位 30 位の比較を行った．検索結果を表 1 に示す.

実験 1-2 IR 分野における安全保障, 経済, 国際機関, 人権問題, 環境問題, 理論, 概念の各サブ領域から IR 専門用語を 15 件選択する.これらを検索語として実験 1-1 と同様に IR 空間において単語検索を行い, 上位 10 位にランクされた単語を正解セットとする．次に，このセットが統合空間において上位 10 位中，上位 20 位中にランクされた割合を計測する.結果は図 2 に示す.

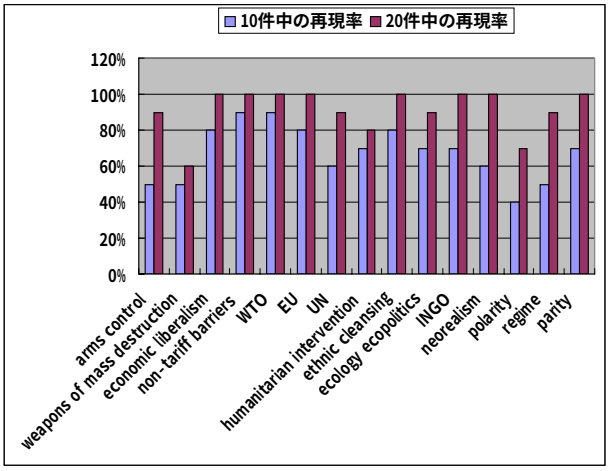

図 2 実験結果 1-2　IR 検索語→IR 用語（統合空間）

実験 1-3 統合空間において，一般語である「trade」「environment」「human」を検索語に設定し，それぞれ上位 10 位以内にランクされる IR 専門用語の割合を計測した．結果を表 2 に示す.

順位	trade	相関量	environment	相関量	human	相関量
1	GATT	0.46484	organization	0.447723	human	0.501031
2	free trade area	0.458932	environment	0.445309	ethnic cleansing	0.4802
3	protectionism	0.447085	pollution	0.403463	genocide	0.430277
4	tariff	0.43101	green movements	0.347933	Genocide Convention	0.415171
5	Tokyo round	0.407401	ecology/ecopolitics	0.308913	immigration	0.367203
6	trade	0.389198	INGO	0.30539	international law	0.35866
7	free trade	0.380862	north/south	0.302781	nationalism	0.352227
8	common market	0.379662	globalization	0.255842	nation	0.329126
9	quota	0.363634	NATO	0.244127	ethnic nationalism	0.328147
10	Kennedy round	0.359655	Earth	0.239894	balkanization	0.325926

表 2　実験 1-3　一般検索語→IR 用語（統合空間）

5.1.2. 実験結果

実験 1-1 IR 空間では「preventive war」「inspection」「second strike」「non-offensive defence」「MAD」といった語が「nuclear weapons（核兵器）」との関連が強いために上位にランクされているが，統合空間を用いた検索では, IR 空間で上位にランクされた単語を残しながらも,「arms control（軍備管理）」に意味的により近い「weapon」「INF treaty」「START II」「START I」「weapons of mass destruction」「cruise missile」などの用語や固有名詞が上位に現れていることが分かる．また,「tactical nuclear weapons」「CTBT」といった「arms control」と意味的に近い語が順位を上げている．この結果は，IR 空間と比較して, 統合空間においてより適切に専門用語の定義づけがなされていることを示している.たとえば「INF treaty」は, IR 空間において「nuclear weapons, arms control, start I, start II nineteen-eighty-nine, Warsaw Pact, bloc, constructive engagement, perception, détente, Gorbachev doctrine, Cold War, inspection」といった関連特徴語のみで特徴づけられていたが，総合空間においてはこれに加え，「arms, weapon, missile, treaty, agreement, remove, zero」といった定義特徴語で特徴づけられている．これらは,「arms control」の総合空間における定義特徴語「arms, control, reduce, remove, weapon...」と共通するため，上位にランクされる結果となったと考えられる.

実験 1-2 IR 空間にて上位 10 にランクされた単語は, 統合空間においても半数以上が 70%の割合で 10 位以内にランクされている．さらに, 上位 20 以内にランクされる割合は，80%～100%の高い数値を示している．たとえば，検索語「economic liberalism（経済リベラリズム）」に対して IR 空間においては「protectionism, GATT, free trade, common market, economic liberalism, tariff, free trade area, Tokyo Round, WTO, quota」が上位 10 位にランクされたが，総合空間では「common market」「WTO」の代わりに「Kennedy Round」「non-tariff barrier」が上位 10 位にランクされている．しかし，統合空間における上位 20 件を見ると，「common market」「WTO」「Kennedy Round」「non-tariff barrier」のいずれもランクされている．これは, 統合空間が IR 空間の基本的な構造を崩していな

いことを示している．

実験 1-3 統合空間では，一般語による検索から関連する専門用語が上位にランクされている．また，上位にランクされた一般語についても，IR 分野の共有知識を反映した語が現れている．これは，一般語についての専門用語による関連づけが適切になされていることを示している．たとえば，「environment」という一般語による検索語からは，「pollution」「green movements」「ecology」「INGO」「globalization」といった地球環境問題に関連が強い IR 用語が上位にランクされている．これは，いずれの単語も総合空間においては「air, water, green, people, plant, Earth」といった「environment」の定義特徴語と共通する定義特徴語で特徴づけられているためと考えられる．

5.1.3. 考察

以上の結果から，IR 空間のみに比べ，統合空間のほうがより適切に単語の定義づけがなされていることが検証された．また，統合空間を用いると，一般語と IR 専門用語の双方から，IR 専門用語のみならず IR 分野の意味を反映させた一般語をも検索可能であることが検証された．

5.2. 実験 2

4.2 のプロセスを経て生成された統合空間を用いてドキュメントの意味的連想検索の実験を行う．

5.2.1. 実験方法

インターネット上に存在する IR 関連のドキュメント 40 件を収集し，検索対象ドキュメントとした．ドキュメントに対するメタデータ設定としては，1）IR 専門用語のみ，2）一般語のみ，3）IR 専門用語および一般語の三種類のメタデータを用意した．具体的な手順としては，ドキュメント中に出現した単語から名詞と動詞のみを抽出し，さらにその中から空間生成に用いた特徴語群に含まれるものを抽出し，これを 2）一般語のみメタデータに設定した．次に，内容から判断して関連する IR 専門用語を追加し，3）IR 専門用語および一般語とした．最後に，3）から一般語を削除し，1）IR 専門用語のみとした．一例として 3）のメタデータ設定を表 3 に示す．また，検索語についても同様，三種類の検索語を用意した．ドキュメントのメタデータ設定と検索語との，3×3 通りの実験の組み合わせを図 3 に示す．

表 3　メタデータ設定例

ID	メタデータ
doc1	trade system, free trade, tariff, regime, import, product, trade1, standard1, success…
doc2	aid, north-south, LDCs, forth-world, poor, people1, aids, die1…
～	～
doc15	human rights, war, intervention, communal conflict, prisoner, race, soldier, rights, attack …
doc16	epistemic communities, futurology, …global governance, future, government, theory, idea…
～	～
doc39	water, natural resources, world-politics, environment , stop, dirty, air, water, protect, earth…
doc40	armistice, war, conflict, demilitalization, intelligenge, stop, attack, fight, information …

図 3　実験の組み合わせ

		検索語		
		IR語のみ	一般語のみ	IR語+一般語
メタデータ	IR語のみ	Ⅰ	Ⅱ	Ⅲ
	一般語のみ	Ⅳ	Ⅴ	Ⅵ
	IR語+一般語	Ⅶ	Ⅷ	Ⅸ

次に，安全保障に関する IR 専門用語「conflict（紛争）」と一般語「attack（攻撃）」を検索語として，内容から両者に共通する安全保障や紛争に関するドキュメント 8 件をあらかじめ正解ドキュメントとし，ID を doc5, doc10, doc15, doc20, doc25, doc30, doc35, doc40 とした．図 3 の 9 通りの組み合わせに従い検索を行った結果を，表 4 に示す．

表 4　実験 2　統合空間における検索結果

		検索語：conflict		検索語: attack		検索語：attack , conflict	
	順位	ID	相関量	ID	相関量	ID	相関量
IR語のみメタデータ →	1	doc40	0.502226	doc5	0.501295	doc5	0.544352
	2	doc5	0.492371	doc40	0.392284	doc40	0.535234
	3	doc15	0.431048	doc15	0.355318	doc15	0.45911
	4	doc20	0.337915	doc20	0.284476	doc20	0.343914
	5	doc25	0.268766	doc37	0.202732	doc25	0.256491
	6	doc35	0.246094	doc11	0.201073	doc11	0.242029
	7	doc11	0.241361	doc9	0.199328	doc35	0.236115
	8	doc2	0.233853	doc17	0.199207	doc2	0.235324
	9	doc30	0.231785	doc23	0.195791	doc30	0.233203
	10	doc33	0.220733	doc25	0.193684	doc33	0.199303
一般語のみメタデータ →	1	doc40	0.510731	doc10	0.512244	doc40	0.552874
	2	doc30	0.424205	doc20	0.446176	doc30	0.45841
	3	doc10	0.259057	doc40	0.433106	doc10	0.32358
	4	doc20	0.24301	doc30	0.364944	doc20	0.295688
	5	doc24	0.222061	doc14	0.284089	doc24	0.218491
	6	doc35	0.198955	doc5	0.277443	doc35	0.204756
	7	doc5	0.188389	doc8	0.26781	doc5	0.199559
	8	doc26	0.181802	doc25	0.264337	doc25	0.178692
	9	doc32	0.181352	doc23	0.249578	doc26	0.169355
	10	doc11	0.179863	doc15	0.24773	doc6	0.16754
IR語+一般語メタデータ →	1	doc5	0.472904	doc5	0.493925	doc5	0.52478
	2	doc40	0.468055	doc20	0.394756	doc40	0.498384
	3	doc15	0.390095	doc40	0.378979	doc15	0.416298
	4	doc30	0.342379	doc15	0.340018	doc30	0.3652
	5	doc20	0.311306	doc10	0.322031	doc20	0.341837
	6	doc25	0.264962	doc30	0.299353	doc25	0.258615
	7	doc35	0.255268	doc14	0.271628	doc35	0.250516
	8	doc33	0.222828	doc17	0.254037	doc10	0.23775
	9	doc38	0.219296	doc23	0.245184	doc2	0.209467
	10	doc10	0.21916	doc8	0.236924	doc24	0.207355

5.2.2. 実験結果

IR 語および一般語でメタデータ設定を行い，IR 語および一般語で検索を行った場合（Ⅸ）が最も検索精度が高いことが分かる．また，IR 語のみ，一般語のみでメタデータを設定した場合でも，IR 語および一般語で検索をかけると再現率が高い（Ⅲ，Ⅵ）．さらに，IR 語のみのメタデータ設定で一般語を検索にかけた場合（Ⅱ），および，一般語のみでメタデータ設定を行い IR 語のみで検索をかけた場合（Ⅳ）でも，精度は高くないが，少なくとも 5 件は上位 10 位にランクインされていることが分かる．

5.2.3. 考察

以上の結果から，統合空間を用いると，一般語の検索語から IR 用語のみでメタデータ設定をしたドキュメントが検索可能であること，逆に，IR 用語の検索語から一般語のみでメタデータ設定をしたドキュメントを検索できることが検証された．また，メタデータ設

定および検索語選択において, IR 用語と一般語を併用すると最も検索精度が高いことが分かった.

5.3. 実験全体の考察

実験 1 および実験 2 より, IR 基本空間と比較して統合空間のほうがより適切に文脈を反映していることが分かる. また, 統合空間を用いると, 一般語の検索語から専門用語を用いたドキュメントが検索可能となり, 専門用語の検索語から一般語を用いたドキュメントが検索できることが検証された.

6. 結論：本方式の意義と可能性

本稿では, 国際関係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生成方式を示した. 本方式により, 一般語辞書をもとに構築された既存のベクトル空間の行列に国際関係用語集をもとに構築した行列を合成し, 新たな統合空間を生成することが実現可能となることを示した. また, 本方式を意味的連想検索へ適用し, 国際関係分野を対象とした検索システムの実現ならびに実験により, その実現可能性と精度を検証した.

本方式により生成した意味的連想検索空間を用いることにより, 国際関係分野の言葉やドキュメントといったデータ間の意味的な関係を相関量として計算し, 意味的に関連する言葉およびドキュメントを検索することが可能となる. 特に, 時間的・空間的に幅広い内容を含む国際関係分野のドキュメントについて, 分析者の視点や分析時の文脈に応じてダイナミックに言葉の意味を扱う動的クラスタリングや意味的データマイニング [13]の応用可能性は高いと考えられる. 様々な年代のドキュメント群について, 時系列的な変化, たとえば政策決定者の認知や態度の変化を分析すること, あるいは, 相手アクター（国・組織・政府・国民）の違いによる政府関係者の発言内容の違いを分析すること, 公式文書と非公式文書の意味内容の違いを分析すること等, 様々な応用可能性が考えられる. 特に今後は, 動的クラスタリングを時系列ドキュメント分析に応用し, ある概念の使われる文脈が時系列的にどのように変化するかを分析する方式の実現を目指す.

謝辞

本研究の実験にあたり, 多くのご助言を頂いた吉田尚史氏（慶應義塾大学政策・メディア研究科）に感謝いたします.

文　献

[1] Dina A. Zinnes, “A comparison of Hostile Behavior of Decision-Makers in Simulated historical Data,” *World Politics* 18, 1966, pp474-502.

[2] Ole R. Holsti, “Content Analysis,” Gardner Lindzey and Elliot Aronson eds., *The Handbook of Social Psychology,* 1968, pp.596-632.

[3] Holsti and Robert C. North, “Comparative Data from Content Analysis: Perception of History and Economic Variables in the 1914 Crisis,” Richard L. Merritt and Stein Rokkan eds., *Comparing Nations: The Use of Quantitative Data in Cross-National Research*, 1966, pp.169-190.

[4] 猪口孝『国際関係の数量分析-北京・平壌・モスクワ,1961-1966 年』巌南堂, 1970.

[5] Robert Axelrod ed., *The Structure of Decision : The Cognitive Maps of Political Elites*, Princeton U. P., 1976.

[6] 山本吉宣・谷明良「認知構造図」『オペレーションズ・リサーチ』, 1979.

[7] Christer Jonsson ed., *Cognitive Dynamics and International Politics*, London : Frances Printer, 1982.

[8] 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.

[9] 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.

[10] Kiyoki, Y., Kitagawa, T. and Hitomi, Y.：A fundamental framework for realizing semantic interoperability in a multidatabase environment, *Journal of Integrated Computer-Aided Engineering*, Vol.2, No.1, pp.3-20, John Wiley & Sons, Jan. 1995

[11] 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.

[12] 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.15-28, 1999.

[13] 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.

[14] 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, Vol. 43, No.SIG5(TOD14), pp.37-53, 2002.

[15] Evans, Graham and Newnham, Jeffrey : Dictionary of International Relations, Penguin Books, 1998.

[16] Longman Dictionary of Contemporary English, Longman, 1987.