異分野データベース群の意味的検索空間生成・統合プロセスの実現

鷹野孝典† 清木康††

本稿では，異分野データベース群を対象とした意味的検索空間生成・統合プロセスの実現について述べる．異分野のデータベースを連携して，目的のデータを獲得するシステムの実現方式として，異分野データベースを対象とした意味的検索空間統合方式が提案されている．この方式は，分野別に構築された既存のベクトル空間のマトリクスを対象として，分野間の共通概念 (共通語) を用いてマトリクスを統合し，意味の解釈を伴った検索空間の統合を実現する方式である．本稿では，この統合方式に，意味的検索空間を専門知識により修正するプロセスを追加し，意味的検索空間の構築から統合までを系統的に行う方法の実現について述べる．本稿では，エネルギー分野と生活環境分野を対象としたシステムの実現ならびに実験により，提案方式の有効性を確認する．

A Generation and Integration Process for Semantic Retrieval Spaces in a Heterogenious Database Environment

Kosuke Takano† and Yasushi Kiyoki††

We have presented an integration method for semantic retrieval spaces of heterogeneous fields. This method makes it possible to integrate semantic retrieval spaces with the interpretation of meanings by using common concepts(common terms) for matrices of heterogeneous fields. In this paper, we present a process for constructing and integrating semantic retrieval spaces systematically. This process includes functions for modifying semantic retrieval spaces. We show several experimental results for database retrieval in the energy and life environmental fields to clarify the effectiveness of the process and its functions.

1. はじめに

インターネット上には，膨大な情報資源がある．Web ブラウザなどのアプリケーションを通じて，WWW のサーチエンジンや学術研究のための専門分野データベースを利用し，これらの情報資源にアクセスできる．これらのシステムで，広く利用されているデータ検索方式として，パターンマッチングによる方式がある．これに対し，パターンマッチングによらない，意味的なデータ検索方式として Latent Semantic Indexing[1),2)] や意味の数学モデルによる意味的検索方式[3)~5),7),8)] が提案されている．

一方，インターネット上から，利用者が目的とするデータを効率的に獲得するために，複数のデータベースを連携して，データ検索を行うシステムの実現は重要な課題である．このような研究として，メタサーチエンジン[11)] や意味的検索空間統合方式[6),9),10)] が提案されている．メタサーチエンジンは，利用者の問合せに対して有効なデータを，検索可能と考えられる複数の特定データベースに対して問合せを行い，その検索結果を統合して，利用者への検索結果とするシステムである．しかし，さまざまな研究分野において、分野の枠組は決まっていても，その内容は複数分野にまたがるものである．よって，静的に区分された枠組みを超えて，データの相関を計量できるデータ検索システムの実現は重要である．意味的検索空間統合方式は，異種の分野について，それぞれの分野特有の用語の意味的関係を記述したマトリクスを，共通の概念を用いて統合する方式である．この方式を，意味の数学モデルなどの意味的検索方式に適用することで，このような分野の枠組みを超えた，データの相関が計量可能となる．意味的検索空間統合方式を適用した意味的検索システムの有効性として，統合後の意味的検索空間を用いた検索では，単独の意味的検索空間を用いた検索よりも，分野横断的な内容のデータがより上位に検索できることが示されている[6),9),10)]．

本稿では，この意味的検索空間統合方式について，意味的検索空間を専門知識により修正するプロセスを加えた意味的空間統合プロセスを提案する．この修正プロセスを適用して修正された統合意味的検索空間を用いた検索では，修正プロセスを適用しない場合の統合意味的検索空間を用いた検索と比較して，より多くのデータについて，単独の意味的検索空間を用いた検索よりも上位に検索できるようになる．本稿では，実際にエネルギー分野と生活環境分野を対象とした実験システムを構築し，その有効性を検証する．

2. 意味的検索空間生成・統合プロセス

本稿で示す意味的検索空間生成・統合プロセスは，3 章で具体的に述べる意味的検索空間統合方式に，意味的検索空間を専門知識により修正するプロセスを追加し，意味的検索空間の構築から統合までを系統的に行うことを示す方式である．図 1の示す，意味的検索空間統合プロセスの内容について，以下に具体的に述べる．

Process1-1　分野ごとのメタデータ空間の作成

3.2 節で具体的に述べるメタデータ空間作成方式に従い．専門辞書・百科事典などの専門知識を利用して，分野ご

† 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

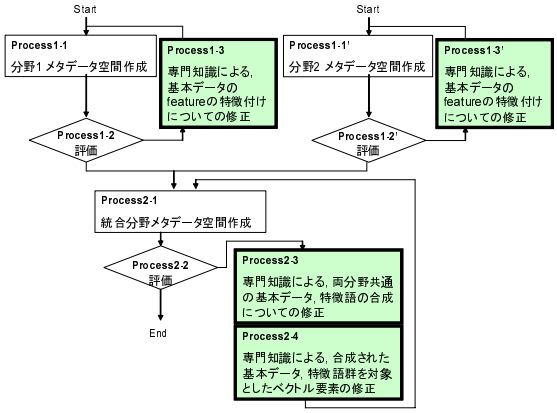

図 1 意味的検索空間生成・統合プロセスの実現

とに独立にメタデータ空間を作成する.

Process1-2 メタデータ空間の評価

メタデータ空間を統合する前提として, 各々のメタデータ空間を用いた検索の検索精度が高いことの検証を行う.

Process1-3 専門知識による, 基本データの feature の特徴付けについての修正

Process1-2 において, 検索精度が基準となる評価を満たさなかった場合, Process1-1 で生成したメタデータ空間にて feature の特徴付けが正当でない基本データが存在する可能性が考えられる. Process1-3 では, そのような特徴付けが正当でない基本データを対象として, 専門知識に基づいた特徴付けの修正を行う.

Process2-1 統合分野メタデータ空間の作成

3.3 で具体的に述べるメタデータ空間統合方式に従い, Process1-1 ～ Process1-3 の手順により生成されたメタデータ空間の統合を行う.

Process2-2 統合分野メタデータ空間の評価

統合分野メタデータ空間を用いた検索の有効性の評価として, Process1-1 ～ Process1-3 で生成されたメタデータ空間では上位に検索されなかった, 分野横断的な内容のメディアデータや幅広く言及している内容のメディアデータが, 統合分野メタデータ空間を用いた検索では, 上位に検索されていることの検証を行う. また, 統合分野メタデータ空間を用いた場合の検索精度について, 単独のメタデータ空間を用いた場合の検索精度と比較し, 評価を行う.

Process2-3 専門知識による, 両分野共通の基本データ, 特徴語の合成についての修正

Process2-2 での統合分野メタデータ空間を用いた検索の評価において, 分野横断的な内容のメディアデータや幅広く言及している内容のメディアデータが上位に検索できていない場合, ないしは統合前の各々のメタデータ空間を用いた検索よりも検索精度が悪い場合に, 分野間の共通概念として合成した基本データ群や特徴語群の中に, 両分野で意味的に等価でないが合成を行った基本データや特徴語が存在する可能性が考えられる. Process2-3 では, 専門知識によりそのような基本データや特徴語を検出した場合, それらの基本データや特徴語に対して, 合成を行わないように設定を行う.

Process2-4 専門知識による, 合成された基本データ, 特徴語群を対象としたベクトル要素の修正

Process2-3 と同様に, Process2-2 での統合分野メタデータ空間を用いた検索において基準となる評価を満たさない場合, メタデータ空間統合方式に従って, 合成された基本データの中に, 両分野で共有できない, ないしは両分野にとって不必要である feature で特徴付けされた基本データが存在する可能性がある. Process2-4 では, このような基本データ, 特徴語群を対象として, 専門知識に基づいたベクトル要素の修正を行う.

3. 意味的検索空間統合方式の意味の数学モデルへの応用と実現への適用

ここでは、すでに提案されている意味の数学モデル[3)~5)], および意味的空間統合方式[6),9),10)] について示し、本稿で提案する空間生成・統合プロセスを意味の数学モデルおよび意味的空間統合方式へ適用する場合の実現方式を示す。

一般的な検索手法であるパターンマッチング検索では，静的かつ明示的に与えられた記述に対する単純なパターン照合でのみ検索を行うが，実際は，データのもつ意味や，データ間の関係性は静的に決め得るものではなく，文脈や状況，あるいはユーザの視点に応じて動的に変化するものである．意味的連想検索では，分野別の専門知識を利用して，その分野の「意味」を形式的に計量することのできるベクトル空間である「メタデータ空間」を生成する．メタデータ空間における文脈解釈，ベクトル計算により，指定した文脈に対して，意味的に近い情報を動的に検索することを可能にしている．

意味的検索空間統合方式を，意味的連想検索に適用することにより，分野別に生成されたメタデータ空間が，意味の解釈を伴って統合される．これにより，分野統合的な意味計量が可能な，意味的連想検索を実現する．

以下に，意味的連想検索に関する概要を述べ，さらに，意味的検索空間統合方式を意味的連想検索に適用する際の実現方法について具体的に述べる．

3.1 意味的連想検索の概要

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，すでに提案されている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[3)~5),7),8)].

(1) **メタデータ空間 $\mathcal{MDS}$ の設定**

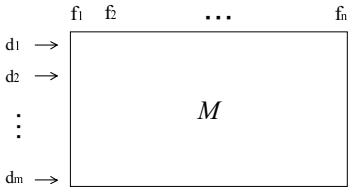

図 2 データ行列 $\mathcal{M}$ によるメタデータ表現

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． $\mathcal{MDS}$ を生成するための基本データ行列 M を図 2 に示す $(f_j : feature\text{-}j,\ d_i : basic\text{-}word\text{-}i)$.

(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．

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

各基本データは，ベクトル表現された特徴を持っている．

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

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 $\mathcal{MDS}$ 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．

(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．

3.2　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[3),10)]．

(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．

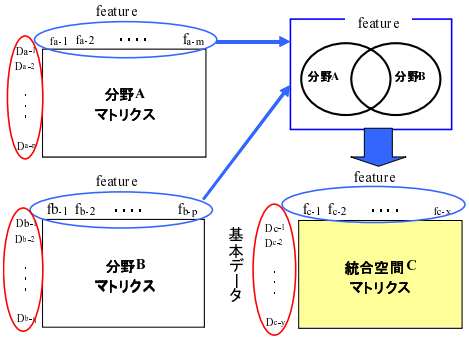

図 3　feature, 基本データの統合

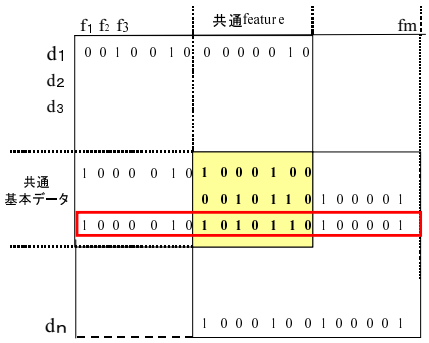

図 4　特徴付け要素の合成

(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，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．

3.3　メタデータ空間統合方式

分野別に生成されたメタデータ空間を対象に，提案方式を適用し，メタデータ空間の意味的統合を実現する[6),9),10)]．ここでは，分野別に生成されたメタデータ空間 A，B を対象に，メタデータ空間を統合し，統合空間 C を実現する際の，具体的なプロセスを示す．

Step-I　A, B間における feature 群の統合

A，B 間において，それぞれの feature 群を合成し，feature 語の重複を除く．この集合を，統合空間 C の feature 群と定義する．このプロセスの概要を図 3に示す．

Step-II　A, B間における基本データ群の統合

同様に，A，B 間において，それぞれの基本データ群を合成し，語の重複を除く．この集合を，統合空間 C の基本データ群と定義する．図 3に示す．

Step-III　要素の統合

A，B それぞれのマトリクスにおいて定義されている 1，-1，0 の各要素を合成する．図 4に要素の合成方法を示す．

A，B の合成マトリクスにおける feature および基本

データの共通部分をαとし，非共通部分をβとする．このプロセスにおいて重要な点は，共通部分αにおいて，A，Bの共通基本データに対する特徴付け設定のオペレーションである．ここでは，A，Bの合成マトリクスにおけるそれぞれの要素について，論理和をとる方法を示す．A，Bの合成マトリクスにおける共通基本データ（D）が，A，Bそれぞれのマトリクスにおいて，feature語f1～f8によって，以下のように特徴付けされているとする

Aにおける D ： f1, f2, f5, f6, f7
Bにおける D ： f1, f3, f4, f6, f8

この場合，共通基本データDに対する特徴付けは，論理和をとって，

Cにおける D ： f1, f2, f3, f4, f5, f6, f7, f8

と特徴付け直す．

また，非共通部分βにおいては，元のマトリクスにおける特徴付け要素をそのまま用いることとする．共通部分αにおける用語の合成方法としては，他にも，論理積をとる，それぞれの要素の値を足す，などいくつかのオペレーションが考えられる．

以上のプロセスにより，メタデータ空間A, Bを統合し，新たな統合メタデータ空間Cが生成される．

3.4 空間生成・統合プロセスのエネルギー，生活環境分野への適用

ここでは，本稿で提案する意味空間生成・統合プロセスを，3.1～3.3章で示した意味の数学モデルおよび意味的空間統合方式へ適用し，エネルギー・生活環境両分野を対象として, 実現する方法について述べる．

具体的には，2章で示した意味空間生成・統合プロセスに従い，次の4つのメタデータ空間を生成する方法について述べる．ここで，エネルギー分野とは，化石燃料, 新エネルギー，エネルギー政策などのトピックを対象とした分野であり，生活環境分野とは，都市, 水資源, 大気汚染, 公害などのトピックを対象とした分野である．

Space-1: エネルギー分野メタデータ空間
Space-2: 生活環境分野メタデータ空間
Space-3: 統合分野メタデータ空間
Space-4: 修正後の統合分野メタデータ空間

2章で示したProcess1-1～Process2-4に従って，これら4つのメタデータ空間Space-1～Space-4を生成する方法について述べる．

Process1-1 エネルギー，生活環境分野メタデータ空間の生成
分野ごとに個別に構築され, それぞれ独立に検索可能であるメタデータ空間として，エネルギー分野を対象とした「エネルギー分野メタデータ空間 [Space-1]」および，生活環境分野を対象とした「生活環境分野メタデータ空間 [Space-2]」を生成する．

Process1-2，1-3 単独分野メタデータ空間の評価・修正
エネルギー分野，生活環境分野，各々の単独分野メタデータ空間を用いた検索において，その検索結果が高い検索精度を示す (Process1-2) まで，各々のメタデータ空間に対し, 専門知識による基本データの feature の特徴付けの修正を行い (Process1-3), それぞれの分野のメタデータ空間を構築する．

Process2-1 エネルギー，生活環境分野メタデータ空間の統合

Context: 水素エネルギー

Order	Rising Info.	Result of Integrated Space(Energy & Life Enviroment)		
1	(life) 1<-575 up	970723229	0.603658	[環日本海環境自治体サミット 海洋汚染に国際協
2	(ener) 2<-7 up	980318025	0.583117	[温暖化防止に挑む] 京都会議から３カ月／３
3		980612049	0.556434	[社説] エネルギー 原子力頼みに三つの疑問]
4	(ener) 4<-60 up	990327042	0.554313	[社説] 温暖化防止 さあ排出量削減の本番だ]
5	(life) 5<-61 up	000619112	0.552813	[地域報道２０００] 木質バイオマス 循環型社
6	(life) 6<-547 up	970825036	0.549720	[社説] 変革のデザイン ごみ列島 循環型社会へ
7	(life) 7<-47 up	000820152	0.549533	[マイクロパワー「２１世紀の発電主流に」 ー米
8	(ener) 8<-239 up	990330117	0.544669	[特集] 環境大学エコレッジ（その３止） 大学
9	(ener) 9<-43 up	990815153	0.540296	[おもかじとりかじ] 省エネにも構造改革を＝江
10		990924037	0.538972	[社説] 新エネルギー 高まる燃料電池への期待
11	(life) 11<-28 up	001006076	0.536580	[揺れるエネルギー政策 ２１世紀の選択肢／下 原
12	(life) 12<-108 up	000621071	0.534992	[経済観測] ドイツの脱原発＝幸兵衛]
13	(ener) 13<-124 up	990317100	0.534436	[特集] ２１世紀危機警告委員会（その２止）

図5 実験システム検索結果画面の例

エネルギー分野，生活環境分野，各々の単独分野のメタデータ空間を用いた検索において，その検索精度が高いことを確認したのち，3.3章で示したメタデータ空間統合方式により統合を行い，新たな「エネルギー・環境分野の統合分野メタデータ空間 [Space-3]」を生成する．ここで，2分野のメタデータ空間の統合には，論理和のオペレーションなどを用いることが考えられる[6),9),10)]．

Process2-2～2-4 統合分野メタデータ空間の評価・修正
統合分野メタデータ空間 [Space-3] を用いた検索において基準となる評価を満たさない場合 (Process2-2), 統合分野メタデータ空間 [Space-3] に対し, 専門知識による, 両分野共通の基本データ，特徴語の合成およびベクトル要素についてのチェック, および修正を行い (Process2-3,2-4)，最終的な統合分野メタデータ空間 [Space-4] を構築する．

以上のプロセスにより，エネルギー分野，生活環境分野，およびこれらの統合分野を対象とした4つのメタデータ空間Space-1～Space-4が生成される．

4. 実　　験

本稿で提案する意味空間生成・統合プロセスを，意味の数学モデルおよび意味的空間統合方式へ適用し，エネルギー・生活環境両分野を対象として，実験システムを実現した．本章では実験結果を示し, その有効性を検証する．

4.1 実験環境

今回の実験では，3.4章で示した4つのメタデータ空間 (Space-1～Space-4) を生成した．エネルギー・生活環境分野の2分野のメタデータ空間の統合には，論理和のオペレーションを用いた．実現した2つの単独メタデータ空間，ならびに統合プロセスを経て生成された統合メタデータ空間のマトリクス構成は，それぞれ表1のとおりである．

なお, 本実験で作成したメタデータ空間では, Process2-3のチェックで検出された合成された基本データや特徴語について, 両分野で意味的に等価でない基本データや特徴語は存在しなかった．また，Process1-3，Process2-4におけるメタデータ空間の修正に関しては，人間が専門知識に基づいて修正を行った．メタデータ空間生成における feature 抽出，および基本データのベクトル特徴付けには，エネルギー分野, 環境分野の専門辞書[12)～20)]を利用した．

表 1 実験システムの詳細

	feature 数	基本データ数	空間次元数
エネルギー分野 メタデータ空間	312	316	302
生活環境分野 メタデータ空間	538	709	538
統合分野 メタデータ空間	667	953	667
共通 (重複) 用語数	183	72	—

表 2 検索対象ドキュメント (Docset-1)

ドキュメントの分野	ID	件数
生活環境分野	doc0xy*	53 件
エネルギー分野	doc1xy*	53 件
両分野共通	doc2xy*	7 件
計	—	113 件

* xy は 2 桁の数字

表 3 メタデータ設定例 (Docset-1)

ID	メタデータ
doc009	水環境 水質汚濁 水力発電 水産業 水資源開発
doc041	ホームオートメーション ライフスタイル 住環境
doc062	プラスチックゴミ 使い捨て文化 ゴミ減量 容器包装
・	・
doc111	原油 重油 パイプライン 石油コークス
doc112	MHD発電 複合発電 燃料電池
doc129	ソフトエネルギー バイオマス 太陽光発電 風力発電
・	・
doc208	石油化学工業 化石燃料 グリーン燃料 LNG DME 環境アセスメント 大気汚染 環境破壊
・	・

本実験では, 検索対象ドキュメントとして, 2 つのドキュメント群を用いた. 1 つめは, エネルギー分野に関連の深いドキュメント 53 件, 生活環境分野に関連の深いドキュメント 53 件, 両分野に関連の深いドキュメント 7 件からなるドキュメント群である. 以下このドキュメント群を, Docset-1 とする. Docset-1 は, メタデータが 3 〜 10 個程度で構成される仮想ドキュメントとして作成した. Docset-1 のメタデータの設定例を, 表 3に示す. また, Docset-1 については, 正解コレクションを, エネルギー分野に関する問合せ 11 件, 生活環境分野に関する問合せ 11 件の合計 22 件の問合せに対し, 各問合せごとに, 検索対象ドキュメントから関連の強いドキュメントを 5 件ずつ選び, 作成した. 正解コレクションの例を表 4に示す. 検索対象ドキュメントの 2 つめとしては, 実際の新聞記事からエネルギー分野に関連の深いドキュメント 575 件, 生活環境分野に関連の深いドキュメント 575 件を抽出したドキュメント群を用いた[21]. 以下このドキュメント群を, Docset-2 とする. Docset-2 は, 新聞記事の本文中に基本データが含まれていた場合, その基本データを, 該当記事のメタデータとして自動設定した. Docset-2 のメタデータ, および記事タイトルの例を表 6に示す.

なお, 本実験では, 実験システムを, Web ブラウザを利用して検索可能なシステムとして実現している (図 5).

4.2 実 験 1

実験 1 では, メタデータ空間統合の前提として, 各々分野別に構築した「エネルギー分野メタデータ空間 (Space-1)」ならびに「生活環境分野メタデータ空間 (Space-2)」それぞれの空間を用いた検索を行い, 個々の検索精度を検証する.

表 4 正解コレクションの例 (Docset-1)

問合せ語	正解ドキュメント
新エネルギー	doc113 doc153 doc154 doc160 doc214
交通機関	doc015 doc031 doc038 doc039 doc040
・	・

表 5 検索対象ドキュメント (Docset-2)

ドキュメントの分野	件数
生活環境分野	575 件
エネルギー分野	575 件
計	1150 件

表 6 メタデータ設定, および記事タイトル例 (Docset-2)

ID	メタデータ	記事タイトル
000224144	ごみ エネルギー ガス プラント メタノール 火力発電 ダイオキシン 原子力発電 固体電解質 焼却 化石燃料 廃棄物 燃料電池 コンビナート 汚染 緑地 ・・・	次世代エネルギー・ 燃料電池の未来／上 排水、ごみも電力に
000106127	エネルギー メタン 車 遺伝子 汚染 環境 森林 再利用 災害 大量生産 新エネルギー 自然 風力発電 ・・・	[環境新世紀] 循環型 社会に向けて／1 自 然との調和を目指し
・	・	・

4.2.1 実 験 方 法

実験データは, エネルギー分野のドキュメント 53 件, 生活環境分野のドキュメント 53 件, 両分野共通のドキュメント 7 件の計 113 件のドキュメント群 (Docset-1) を用いた. 実験では, エネルギー分野・生活環境分野それぞれの検索空間 (Space-1, Space-2) に対して, 各々 11 の問い合わせを発行した. この際, あらかじめ正解とするドキュメントを, 各問い合わせに対して 5 件ずつ設定しておく. 実験結果において, この正解ドキュメント 5 件のうち上位 15 件中, 上位 10 件中に含まれる割合をそれぞれ算出した. この値が高いほど, 望ましいドキュメントが上位に検索されていることを示す.

4.2.2 実 験 結 果

エネルギー分野空間 (Space-1)・生活環境分野空間 (Space-2) を用いた検索結果を, それぞれ図 6, 図 7に示す.

4.2.3 実 験 考 察

上位 15 件中の検索結果において, Space-1 では約 7 割, Space-2 では約 8 割の問い合せで, 0.8 〜 1.0 の高い再現率を示した ([平均再現率]Space-1： 0.75, Space- 2： 0.84). また, 上位 10 件中の検索結果において, Space-1 では約 5 割, Sp ace-2 では 約 6 割の問い合せで, 0.8 〜 1.0 の高い再現率を示した ([平均再現率]Space-1： 0.66, Space-2： 0.78). 本実験により, エネルギー分野, 生活環境分野の各々のメタデータ空間での検索において, 高い精度をもって検索が実現されていることを確認した. 統合分野メタデータ空間を用いた検索における検索精度検証の前提として, 統合の対象となる単独分野メータデータ空間を用いた検索において, 高い検索精度が得られていることを確認した.

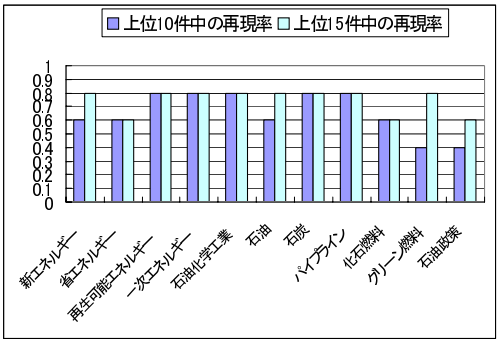

図 6 エネルギー分野メタデータ空間の検索結果

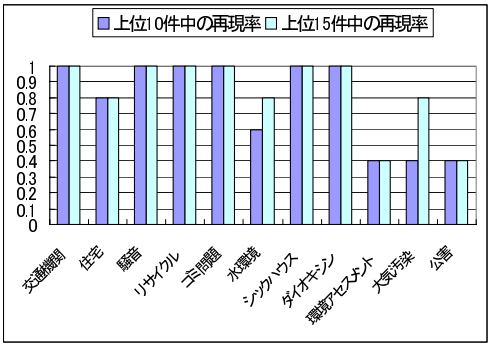

図 7 生活環境分野メタデータ空間の検索結果

4.3 実 験 2

実験 2 では，まず, 2 章で示した意味的検索空間プロセスに従って生成した統合分野メタデータ空間 (Space-4) を用いた検索の有効性を検証する．この検証では，まず, 単独分野メタデータ空間 (Space-1, Space-2) を用いた検索では上位に検索されないような，分野横断的に言及しているドキュメントが，統合分野メタデータ空間 (Space-4) を用いた検索においては上位に獲得可能であることを検証する．次に, 2 章で示した Process2-3, 2-4 に従って修正した統合分野メタデータ空間 (Space-4) を用いた検索では, 修正のない統合分野メタデータ空間 (Space-3) と比較して, より多くの正解ドキュメントについて, 上位に検索可能であることを示す.

4.3.1 実 験 方 法

実験データは, エネルギー分野のドキュメント 53 件, 生活環境分野のドキュメント 53 件, 両分野共通のドキュメント 7 件の計 113 件のドキュメント群 (Docset-1) を用いた．本実験では，まず, 単独分野メタデータ空間 (Space-1 および Space-2) と統合分野メタデータ空間 (Space-3) を用いた検索結果と, 単独分野メタデータ空間 (Space-1 および Space-2) と修正後の統合分野メタデータ空間 (Space-4) を用いた検索結果を比較する．具体的には, Space-1, Space-2, Space-3, Space-4 の 4 つのメタデータ空間に対して, 同一の問合せを発行し, 単独分野メタデータ空間を用いた検索では上位に検索できなかったが，統合分野メタデータ空間を用いた検索では上位に検索されている正解ドキュメントを対象に，その正当性を検証する．次に, 分野横断的な内容のドキュメントについて, 修正後の統合分野メタデータ空間 (Space-4) では, 統合分野メタデータ空間 (Space-3) と比較して, より多くの正解データについて, 上位に検索可能であることを, (1) 順位上昇した正解ドキュメント数, (2) 順位の上昇した正解ドキュメントの平均順位差, (3) 順位上昇後の平均順位, (4) 順位が 6 位以上上昇した正解ドキュメント数, (5) 順位が 6 位以上上昇し, かつ順位上昇後の順位が 10 以内である正解ドキュメント数, の 5 つの視点から比較して検証する.

表 7 「公害」に対する検索結果

doc ID	修正後の統合空間 (Space-4)	修正のない統合空間 (Space-3)	生活環境空間 (Space-2)
doc207	5 位	6 位	46 位
doc208	6 位	7 位	21 位

表 8 「新エネルギー」に対する検索結果

doc ID	修正後の統合空間 (Space-4)	修正のない統合空間 (Space-3)	エネルギー空間 (Space-1)
doc113	8 位	9 位	40 位

表 9 正解ドキュメントの順位上昇に関する修正後の統合空間 (Space-4) と修正のない統合空間 (Space-3) の比較

	修正後 (Space-4)	修正なし (Space-3)
(1) 順位の上昇した正解ドキュメント数	37 個	35 個
(2) 順位の上昇した正解ドキュメントの平均順位差	6.5 位	6.5 位
(3) 順位上昇後の平均順位	7.9 位	7.2 位
(4) 順位が 6 位以上上昇した正解ドキュメント数	12 個	9 個
(5) 順位が 6 位以上上昇し, かつ順位上昇後の順位が 10 以内である正解ドキュメント数	10 個	7 個

4.3.2 実 験 結 果

実験結果を表 7, 表 8, 表 9 に示す.

4.3.3 実 験 考 察

表 7 に示す実験結果では，「公害」に対する検索結果として, 生活環境分野メタデータ空間 (Space-2) における検索では 46 位, 21 位と上位に検索することができなかった正解ドキュメント doc207, doc208 について, 統合分野メタデータ空間 (Space-3) を用いた検索では 6 位, 7 位と上位に検索され, 修正後の統合分野メタデータ空間 (Space-4) を用いた検索では，さらに 5 位, 6 位と上位に検索されていることを示している.

doc208 のメタデータは，「石油化学工業 化石燃料 グリーン燃料 LNG DME 環境アセスメント大気汚染 環境破壊」であり，「石油化学工業 化石燃料」などの，エネルギー分野において “ 公害 “ に関連の強いメタデータに加え，「大気汚染 環境破壊」といった，生活環境の分野において ” 公害 “ に関連の強いメタデータをあわせ持ったドキュメントであることがわかる．このように，分野横断的に広く言及している doc208 のようなデータが，生活環境分野メタデータ空間 (Space-2) を用いた検索では上位に検索されなかったが，修正後の統合分野メタデータ空間 (Space-4) を用いた検索では上位となったことが確認できる．

また, 表 8 に示す実験結果では，「新エネルギー」に対する検索結果として, エネルギー分野メタデータ空間 (Space-1) における検索では 40 位と上位に検索することができなかった正解ドキュメント doc113 について, 統合分野メタデータ空間 (Space-3) での検索では, 9 位と上位に検索され, 修正後の統合分野メタデータ空間 (Space-4) では，さらに 8 位と上

位に検索されていることを示している．doc113 のメタデータ，「アルコール燃料 混合燃料 燃料電池 メタノール自動車 アルコール自動車」といった幅広いメタデータをもっており，doc113 は，より広い視点から述べているドキュメントであることが分かる．このように，総合的なメタデータをもつ doc113 のようなデータは，エネルギー分野のメタデータ空間 (Space-1) を用いた検索では上位に検索されないが，修正後の統合分野のメタデータ空間 (Space-4) を用いた検索では上位に検索できることを確認した．

表 9では修正後の統合分野メタデータ空間 (Space-4) と修正のない統合分野メタデータ空間 (Space-3) を比較して, 単独メタデータ空間より順位の上昇した正解ドキュメントを対象として, 4.3.1 で述べた 5 つの視点について算出した結果を示している．(1) 順位上昇した正解ドキュメント数は, Space-4 を用いた検索では 37 個と, Space-3 を用いた検索の 35 個よりも多くなった．(2) 上昇順位差の平均, および (3) 上昇後の順位平均については, Space-4 と Space-3 で差はほとんどなかった．順位上昇した正解ドキュメント数については, 特に, (4) 順位が 6 位以上上昇した正解ドキュメント数は Space-4 を用いた検索では, Space-3 を用いた検索と比較して, 9 個から 12 個に増え, (5) 順位が 6 位以上上昇しかつ順位 10 位以内である正解ドキュメント数は Space-4 を用いた検索では, Space-3 を用いた検索と比較して, 7 個から 10 個に増えた．これらの結果から, 修正後の統合分野メタデータ空間 (Space-4) を用いた検索では, 修正のない統合空間 (Space-3) よりも特に上位に順位上昇した正解ドキュメント数が増えていることが確認できる．

以上の結果から，エネルギー分野・生活環境分野といった単独分野のメタデータ空間を用いた検索では上位に検索することができなかった，分野横断的なドキュメントや，総合的に多くのことに言及しているドキュメントが，2 章で示した意味的検索空間統合プロセスにしたがって生成した統合メタデータ空間を用いた検索においては, 分野統合的な意味解釈による検索によって，上位に検索されることを確認できた．また, 修正後の統合分野メタデータ空間を用いた検索では, 修正のない統合分野メタデータ空間と比較して, より多くの正解ドキュメントについて, 上位に検索可能であることを確認できた．この実験結果は, 2 章で示した「意味的検索空間統合プロセス」の有効性を示している．

4.4 実　験　3

実験 3 では，2 章で示した意味的検索空間プロセスに従って生成した統合分野メタデータ空間 (Space-4) を用いた検索が, ドキュメント件数, および設定したメタデータ数が大規模である検索対象ドキュメント群においても有効であることを検証する．

4.4.1 実 験 方 法

本実験では, 単独分野メタデータ空間 (Space-1 および Space-2) と修正後の統合分野メタデータ空間 (Space-4) を用いた検索結果を比較する．具体的には, Space-1, Space-2, Space-4 の 3 つのメタデータ空間に対して, 同一の問合せを発行し, 単独分野メタデータ空間を用いた検索では上位に検索できなかったが，統合分野メタデータ空間を用いた検索では上位に検索可能であることを, 大規模な検索対象ドキュメント群として新聞記事データ 1150 件 (Docset-2) を用いて検証する．

4.4.2 実 験 結 果

実験結果を表 10, 表 11に示す．

表 10　「太陽光発電」に対する検索結果

doc ID	修正後の統合空間 (Space-4)	エネルギー空間 (Space-1)
980822160	2 位	165 位

表 11　「新エネルギー」に対する検索結果

doc ID	修正後の統合空間 (Space-4)	生活環境空間 (Space-2)
000106127	46 位	215 位

4.4.3 実 験 考 察

表 10 に示す実験結果では，「太陽光発電」に対する検索結果として, 生活環境分野メタデータ空間 (Space-2) における検索では 165 位と上位に検索することができなかったドキュメント 980822160 について, 統合分野メタデータ空統合分野メタデータ空間 (Space-4) を用いた検索では, 2 位と上位に検索されていることを示している．

980822160 は, メタデータが「エネルギー ガス 家電 開発 原子力原子力発電 施設 消費 省エネルギー 新エネルギー 水 水力 水力発電 製品 炭 都市発電 風力 冷暖房」, また記事タイトルが「温暖化防止はエネルギー源の開発より省エネで」と設定されており、主に環境問題である温暖化対策として, 初期に導入すべき技術の一つとして新エネルギーである太陽光や風力によるエネルギー源が提示されている内容のドキュメントである．このように，エネルギーと生活環境の両分野に横断的に言及している 980822160 のようなデータが，エネルギー分野メタデータ空間 (Space-1) を用いた検索では上位に検索されなかったが，統合分野メタデータ空間 (Space-4) を用いた検索では上位に検索可能なことが確認できる．

また, 表 11 に示す実験結果では，「省エネルギー」に対する検索結果として, 生活環境分野メタデータ空間 (Space-1) における検索では 215 位と上位に検索することができなかったドキュメント 000519031 について, 統合分野メタデータ空統合分野メタデータ空間 (Space-4) を用いた検索では, 46 位と上位に検索されていることを示している．

000519031 は, メタデータが「エネルギー 化石燃料 環境 経済 国際 自動車 消費 新エネルギー石炭 石油 炭素税 電気事業 ・・・」, また記事タイトルが「［ニュースキー２０００］炭素税提案 導入、なおハードル――業界は「死活問題」」と設定されており、主に炭素税の導入という, 環境問題とエネルギー産業の話題の両方に関連の深い内容のドキュメントである．このように，エネルギーと生活環境の両分野に関係の深い内容に言及している 000519031 のようなデータが，生活環境分野メタデータ空間 (Space-2) を用いた検索では上位に検索されなかったが，統合分野メタデータ空間 (Space-4) を用いた検索では上位に検索可能なことが確認できる．

以上の結果から，大規模な検索対象ドキュメント群を用いた場合においても, エネルギー分野・生活環境分野といった単独分野のメタデータ空間を用いた検索では上位に検索することができなかった，分野横断的なドキュメントや，両分野に関連の深い内容のドキュメントが，2 章で示した意味的検索空間統合プロセスにしたがって生成した統合メタデータ空間を用いた検索においては, 分野統合的な意味解釈による検索によって，上位に検索されることを確認できた．この実験結果は, 2 章で示した「意味的検索空間統合プロセス」によって統合されたメタデータ空間を用いた検索が, 大規模な

検索対象ドキュメント群を用いた場合においても有効であることを示している.

5. 結　　論

今回の実験結果から, 本稿で提案する意味空間生成・統合プロセスにより作成した統合分野の意味的検索空間では, 分野横断的な内容のドキュメントや広い内容のデータが単独分野の意味的検索空間よりも, 上位に検索可能であることを確認できた. 特に, 統合分野の意味的検索空間に対し, Process2-4 のような, 合成された基本データの中に, 両分野で共有できない, ないしは両分野にとって不必要である feature で特徴付けされた基本データについて, 専門知識に基づいたベクトル要素の修正を行うことで, 修正のない統合分野空間よりも, より多くの正解ドキュメントについて, 上位に検索可能であることを確認した. また, 統合分野の意味的検索空間を用いた検索が, 大規模な検索対象ドキュメント群を用いた場合においても有効であることを確認した. 以上の結果から, 本稿で提案する, 意味的検索空間統合方式に, 意味的検索空間を専門知識により修正するプロセスを追加する方式は, 有効であったと言える.

今後の課題として, 実験システムの検索結果の評価をより解析的に行うことが考えられる. また, 今回は統合を行う分野として, エネルギー分野, 生活環境分野を設定した. 今後は, 情報通信学, 生命科学, 国際関係学など, より複数の分野の空間を対象とした実験を行い, 本方式の有効性を検証する予定である.

謝辞　本研究に関して, 多くの貴重なご助言を頂いた筑波大学電子・情報工学系北川高嗣教授, 慶應義塾大学 SFC 吉田尚史氏に感謝申し上げます.

参 考 文 献

1) 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).
2) Dumais, S. T., Furnas, G. W., Landauer, T. K., and Deerwester, S.: Using latent semantic analysis to improve information retrieval. In Proceedings of CHI'88: Conference on Human Factors in Computing, New York: ACM, pp. 281–285.
3) 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 (1993).
4) 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).
5) Kiyoki, Y., Kitagawa, T. and Hitomi, Y.: A fundamental framework for realizing semantic interoperability in a multidatabase environment, Journal of Integrated Computer-Aided Engineering, John Wiley & Sons, Vol. 2, No. 1, pp. 3–20 (1995).
6) Y. Kiyoki, and S. Ishihara, : A Semantic Search Space Integration Method for Meta-level Knowledge Acquisition from Heterogeneous Databases, Information Modelling and Knowledge Bases (IOS Press), Vol. 14 (2002).
7) 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).
8) 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, Vol. 40, No. SIG5(TOD2), pp. 15–28 (1999).
9) 石原冴子, 清木康, 吉田尚史：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. データベースと Web 情報システムに関するシンポジウム論文集（Proceedings of DBWeb2001）, Vol. 2001, No. 17, pp. 265–272 (2001).
10) 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. SIG05(TOD14), pp. 37–53 (2002).
11) Meng, W., Yu, C., Liu, K. : Building efficient and effective metasearch engines, ACM Computing Surveys, Vol. 34, No. 1, pp. 48–49 (2002).
12) 日外アソシエーツ編集部. 環境問題記事索引. 日外アソシエーツ (1999).
13) 日本経済新聞社電子メディア局記事データベースグループ. 日経シソーラス. 日本経済新聞社 (1999).
14) ニュース・シソーラス. 中日新聞本社 (1990).
15) 国立国会図書館図書部. 国立国会図書館件名標目表 第5版 (1990).
16) マグローヒル科学技術用語大辞典 改訂第3版 CD-ROM 版. 日刊工業新聞社 (2001).
17) 長倉三郎他. 岩波 理化学辞典 第5版. 岩波書店 (1999).
18) 新村出. 岩波 広辞苑 第5版 CD-ROM 版. 岩波書店 (1998).
19) 現代用語の基礎知識 2001 CD-ROM 版. 自由国民社 (2001).
20) 日外アソシエーツ. DCS- 環境問題情報事典 CD-ROM 版. 日外アソシエーツ (2001).
21) 毎日新聞. CD-5yrs. 毎日新聞 1996-2000. 日外アソシエーツ. (2001).