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

An Integration Process for Semantic Retrieval Spaces in a Heterogenious Database Environment

鷹野 孝典♠ 清木 康♦
Kosuke TAKANO Yasushi KIYOKI

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．また，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．

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[8]や意味的検索空間統合方式[6]，[7]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，検索可能な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，利用者への検索結果とするシステムである．しかし，様々な研究分野において，分野の枠組が決められていても，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，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[2]～[5]に適用することで，分野の枠組を超えた，データの相関が計量可能となる[6]，[7]．

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,分野横断的な内容を持つより多数のデータについて，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，実際にエネルギーと生活環境両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. 意味的検索空間統合のプロセス

本稿で示す意味的検索空間統合プロセスは，3章で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述べる．

Process1-1 分野ごとのメタデータ空間の作成 3.2節で述べるメタデータ空間作成方式に従い，専門辞書等の専門知識を利用し，分野ご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．

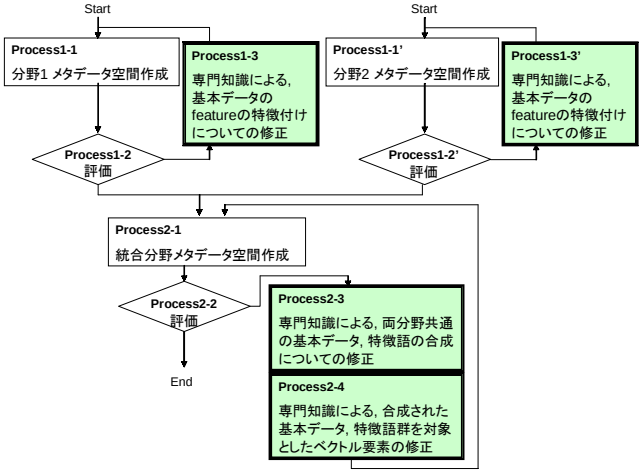

図1 メタデータ空間統合プロセスの実現
Fig.1 Process of integrating metadata spaces

Process2-1 統合分野メタデータ空間の作成 3.3で述べるメタデータ空間統合方式に従い，Process1-1～1-3の手順にて生成されたメタデータ空間の統合を行う．

♠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

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，Process 1-1〜1-3で生成された単独の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，基準とする評価を満たさない場合に，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 Process2-2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的連想検索への応用と実現

意味的連想検索[2]〜[5]および意味的検索空間統合方式[6], [7]の概要を示す． 詳細は[2]〜[7]に示されている．

3.1 意味的連想検索の概要

意味的連想検索方式は，メタデータ空間における文脈解釈，ベクトル計算により，利用者が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とする検索方式である．ここで，メタデータ空間とは，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．

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[2]〜[7].

(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．

(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，否定的な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．

(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．

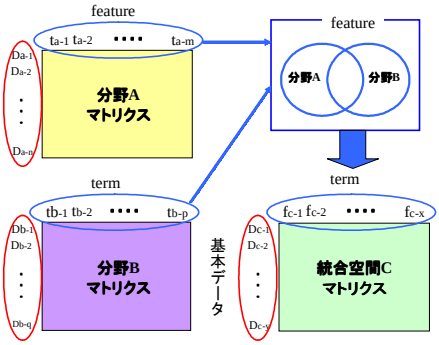

図 3 feature，基本データの統合
Fig.3 Integragion of features and basic data

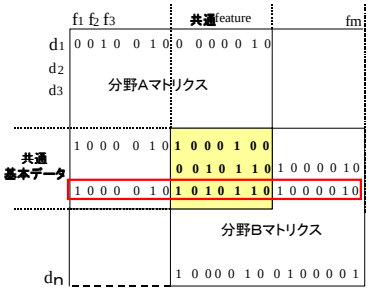

図 4 特徴付け要素の合成
Fig.4 Composition of elements of features

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 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[8]，[9].

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 語の重複を除いた集合を，統合空間 C の feature 群と定義する(図 3).

Step-Ⅱ A,B間における基本データ群の統合 A，B 間において，それぞれの基本データ群を合成し，語の重複を除いた集合を，統合空間 C の基本データ群と定義する(図 3).

Step-Ⅲ 要素の統合 A，B の両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). 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．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するため，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，論理積，それぞれの要素の値を足す，等のオペレーションを行う．論理和のオペレーションについては，[6],[7]に詳細が述べられている．また，非共通部分βにおいては，元のマトリクスの特徴付け要素をそのまま用いる．

4. 実験

4.1 実験環境

Web ブラウザを利用して検索可能な実験システムを構築した．分野ごとの独立なメタデータ空間としては，2 分野を対象とし，化石燃料，新エネルギーなど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」，および都市，公害など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」を設定した．2 章のプロセスに従い，次の 4 つのメタデータ空間を生成した．2 分野のメタデータ空間の統合には，

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

表 1　実装システムの詳細
Table 1 Description of the system

ドキュメントの分類	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

表 2　検索対象ドキュメント(xy は 2 桁の数字)
Table 2 Target documents for retrieval

ID	メタデータ
doc011	ホームオートメーション ライフスタイル 住環境
doc113	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
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

表 3　メタデータ設定例
Table 3 Example of metadata settings

論理和のオペレーションを用いた[6], [7].

Space-1: エネルギー分野メタデータ空間　(Process1-1～1-3)
Space-2: 生活環境分野メタデータ空間　(Process1-1～1-3)
Space-3: 統合分野メタデータ空間　(Process1-1～2-1)
Space-4: 修正後の統合分野メタデータ空間 (Process1-1～2-4)

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，統合分野メタデータ空間にて，Process2-3 にて検出された基本データや特徴語は存在しなかった．また，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[9]～[11]等を利用した．検索対象ドキュメントとしては，エネルギー分野，生活環境分野，両分野それぞれに関連の深いものを選び，表 2 に掲げる件数のドキュメントを用いた．これらのドキュメントは, 3～10 個程度のメタデータが設定されている．メタデータ設定例を表 3 に示す．また，正解コレクションとして，エネルギー分野，生活環境分野に関係する問合せそれぞれ 11 件，合計 22 件の問合せに対し，問合せごとに関連の強いドキュメントを 5 件ずつ選び作成した．

4.2 実験 1

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「エネルギー分野空間(Space-1)ならびに「生活環境分野空間(Space-2)」を用いた検索を行い，検索精度を検証する．

4.2.1 実験方法

実験では，エネルギー・生活環境両分野それぞれの検索空間に対し，各々11 の問い合わせを発行した．この際，あらかじめ正解ドキュメントを，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，上位 15 件中，上位 10 件中に，この正解ドキュメント 5 件が含まれる割合を算出する．

4.2.2 実験結果・考察

エネルギー・生活環境両分野の空間を用いた検索結果を図 5，図 6 に示す．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，Space-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). 本実験より，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タデータ空間(エネルギー分野，生活環境分野)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．

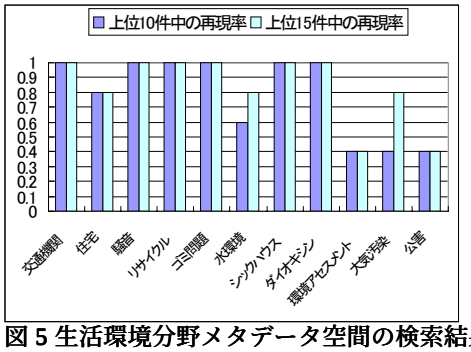

図 5 生活環境分野メタデータ空間の検索結果
Fig.5 Result on metadata space of life environment field

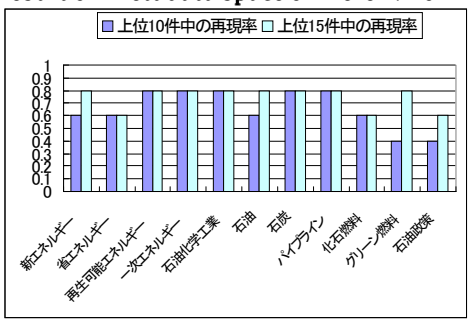

図 6 エンルギー分野メタデータ空間の検索結果
Fig6. Result on metadata space of energy field

4.3 実験 2

2 章の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を用いた検索の有効性を検証する．次に, 2 章の Process2-3，2-4 に従って修正した統合分野メタデータ空間を用いた場合，修正のない統合分野メタデータ空間を用いた場合と比較して，より有効な検索が可能なことを示す．

4.3.1 実験方法

本実験では，Space1～Space4 の 4 つのメタデータ空間，および表 1 に示す計 113 件のドキュメントを実験データとして用いた．まず，単独分野空間(Space-1，Space-2)，統合分野空間(Space-3)，修正後の統合分野空間(Space-4)を用いた検索結果を比較する．具体的には, Space-1～Space-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，修正後の統合分野空間(Space-4)では，統合分野空間(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，表 5 に掲げた 5 つの視点から比較して検証する．

4.3.2 実験結果・考察

実験結果を表 4，表 5，表 6 に示す．表 4 に示す実験結果では，「公害」に対する検索結果として，正解ドキュメント doc208 について，生活環境分野空間(Space-2)における検索では 21 位と上位に検索できなかったが，統合分野空間(Space-3)では 7 位，修正後の統合分野空間(Space-4)を用いた検索では 6 位とさらに上位に検索されている. doc208 は,

docID	修正後の統合空間	修正のない統合空間	生活環境空間

	(Space-4)	(Space-3)	(Space-2)
doc208	6 位	7 位	21 位

表 4 「公害」に対する検索結果
Table 4 Result of “Pollution”

docID	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-2)
doc113	8 位	9 位	40 位

表 5 「新エネルギー」に対する検索結果
Table 5 Result of “New energy”

	修正後 (Space-4)	修正前 (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

表 6 正解ドキュメントの順位上昇に関する比較(Space-3，4)
Table 6 Comparison about rise of ranking of correct answer documents

表 3 に掲げるメタデータが設定されており，“公害“に関連の強いメタデータとして，「石油化学工業 化石燃料」などエネルギー分野の単語に加え，「大気汚染 環境破壊」など生活環境分野の単語が設定されている．表 5 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，正解ドキュメント doc113 について，エネルギー分野空間(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が，統合分野空間(Space-3)での検索では 9 位，修正後の統合分野空間(Space-4)での検索では 8 位とさらに上位に検索されている．doc113 は，表 3 に掲げる両分野共通で関連の深いメタデータが設定されている．このように，分野横断的に広く言及している doc208 のようなデータや，両分野共通で関連のあるメタデータをもつ doc113 のようなデータが，単独分野空間(Space-1，Space-2)を用いた検索では上位に検索されなかったが，修正後の統合分野空間(Space-3，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 6 は修正後の統合分野空間(Space-4)と修正のない統合分野空間(Space-3)を比較して，単独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，5 つの視点について算出した結果を示している．表 6 の(1)，(4)，(5)は順位上昇した正解ドキュメント数についての結果を示している．この結果より，修正後の統合分野空間(Space-4)を用いた検索では，修正のない統合空間(Space-3)よりも，より上位へ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．

5. まとめと今後の課題

今回の実験結果から，2 章で示した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 の専門知識に基づいたベクトル要素の修正を行った場合，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．

以上の結果から，本稿で提案する，意味的検索空間統合プロセスは有効であったと言える．今後の課題として，新聞記事，学術論文等，大規模実験データを対象としてシステム評価を解析的に行うこと，および，情報通信学，生命科学，国際関係額など，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ことが考えられる．

[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 SFC 吉田尚史氏にこの場を借りて感謝申し上げます．

[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] 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).
[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] 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(1995).
[5] 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).
[6] 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).
[7] 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, SIG05(TOD14), pp.37-53(2002).
[8] Meng, W., Yu, C., Liu, K. : “Building efficient and effective metasearch engines”, ACM Computing Surveys, Vol.34, No.1, pp.48-49.(2002).
[9] 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).
[10] 長倉三郎他. 岩波 理化学辞典 第５版. 岩波書店. (1999).
[11] 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).
他.

鷹野 孝典 Kosuke TAKANO
慶應義塾大学政策・メディア研究科修士課程在学中．1998 慶應義塾大学環境情報学部卒業．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．

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．