2009年5月16日発行　第19巻2号　　学術刊行物 ISSN 0917-1436

Journal of Japan Society of Information and Knowledge

情報知識学会誌

Vol.19　No.2　(May 2009)

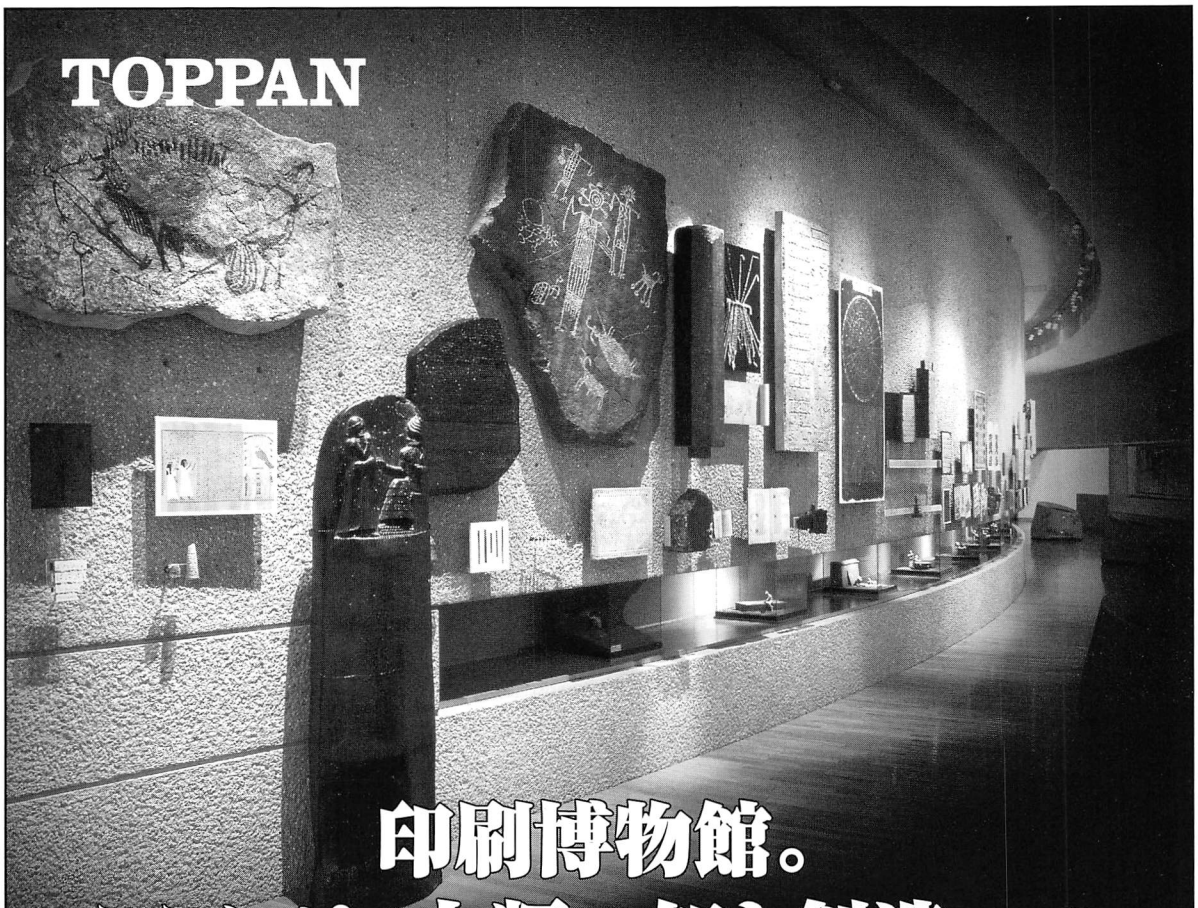
～～～～～～～～～～～～～～～～～～～～目　次～～～～～～～～～～～～～～～～～～～～

特集　第17回（2009年度）年次大会（研究報告会＆総会）

トピックマップを用いた人名典拠情報の構築……研谷 紀夫, 内藤 求 … 57
国内大学図書館におけるデジタルアーカイブの現状……鈴木 良徳, 時実 象一 … 63
博物館における業務情報の共有とIML（Inter-Museum Loan）システムの可能性 ……田良島 哲 … 70
自然言語処理を用いた格言引用システムの開発……牧野 晃典, 梶川 裕矢 … 74
児童書における文章の長さの変化……山本 昭, 日下 陽子 … 80
生物分野におけるWeb APIの適用 ……重元 康昌, 桑名 良和, 權 娟大, 菅原 秀明 … 86
バイオメタデータベースの構築とその応用……鈴木 智典, 宮崎 智 … 92
個々の疾患に特有な関連遺伝子を特定するアルゴリズム……權 娟大, 菅原 秀明 … 98
統合システムを有する分散型熱物性データベースの展開……山下 雄一郎, 馬場 哲也 … 104
マテリアルリスク指標データベースの設計と活用……芳須 弘, 原田 幸明, 藤田 充苗 … 112
テキスト批評の計量化に向けて－書評の計量分析－……村井 源, 徃住 彰文 … 120
村上春樹の初期三部作における構造解析……工藤 彰, 村井 源, 徃住 彰文 … 126
文学作品群の特徴的語彙と概念カテゴリーの抽出 ― 星新一ショートショートの計量分析―
……佐藤 知恵, 村井 源, 徃住 彰文 … 132
音楽評論論文にみる概念構造の変遷 ―ネットワーク中心性を用いた音楽概念の抽出―
……河瀬 彰宏, 村井 源, 徃住 彰文　138
心の状態と言語的特徴：ブログにおける商品紹介文の分析……斉藤 香里, 村井 源, 徃住 彰文 … 144
書評中の語を用いた図書に対する感性パラメータ自動設定の最適化……原田 隆史, 池内 正明 … 152
論文数・引用数からみたわが国の大学間格差の動向分析……根岸 正光 … 158
ネットワーク指標を用いた学際的な論文の抽出……梶川 裕矢, 森 純一郎 … 170
産学連携に関するWeb情報の分析 －大学サイトの事例研究－
……柿沼 澄男, 孫 媛, 西澤 正己, 大山 敬三, 根岸 正光 … 174
論文からの解析手順知識の網羅的抽出から分かる解析技術の経年的変化の分析……荒木 次郎 … 179
協調フィルタリングを用いた共著関係の予測……堀 幸雄, 中山 堯, 今井 慈郎 … 185
世代間共同研究による知識移転のモデル化……福田 毅哉, 梅本 勝博 … 191
完全合理性と限定合理性……安平 哲太郎 … 195
情報過程の諸相……棚次 奎介 … 212
思考展開図を用いたSQL文の可視化手法 ……村川 猛彦, 中川 優 … 218
サーチエンジン検索結果ページにおける視線情報の分析
……高久 雅生, 江草 由佳, 寺井 仁, 齋藤 ひとみ, 三輪 眞木子, 神門 典子 … 224

～～～～～～～～～～～～～～～～～～～

TOPPAN



印刷博物館。
ここには、人類の知と創造への
エネルギーがあふれています。

絵画と文字の始原を求める…。先人たちの知の遺産に触れる…。
そして、印刷とコミュニケーションの過去、現在、未来の姿を探る。
東京・文京区に開館した日本初の本格的な「印刷博物館」。
ここは人類の偉大なる知と創造へのエネルギーを感じることができるスペースです。



printing museum, Tokyo
印刷博物館

〒112-8531
東京都文京区水道1丁目3番3号
TEL：03-5840-2300（代）
http://www.printing-museum.org/

●交通：JRおよび地下鉄有楽町線、東西線、南北線、大江戸線飯田橋駅より徒歩約13分。地下鉄有楽町線江戸川橋駅より徒歩約8分。地下鉄丸の内線、南北線後楽園駅より徒歩約10分。●開館時間：10時～18時（入場は17時30分まで）●休館日：毎週月曜日（但し祝日の場合は翌日）、年末年始、展示替え期間●入館料：一般（中学生以上）300円、小学生100円、団体割引あり（税込）

情報知識学会 第17回（2009年度）年次大会（研究報告会および総会）

実行委員長　研谷紀夫（東京大学）
副委員長　江草由佳（国立教育政策研究所）
副委員長　高久雅生（物質・材料研究機構）
副委員長　村井源（東京工業大学）

1. 日時：2009年5月16日（土）
2. 会場：東京工業大学 大岡山キャンパス 西9号館（東京都目黒区大岡山２丁目１２-１）
（最寄り駅: 大岡山駅より徒歩1分）
3. 参加費：無料
4. 資料代：会員無料、一般非会員3,000円、学生非会員1,000円
5. 情報交流会参加費：1,000円（学生無料）

プログラム

時間	A会場（W935）	B会場（W934）
9:00	（受付開始）	
9:20〜10:00	総会	—
	セッション A-1 （座長: 田窪 直規）	セッション B-1 （座長: 国沢 隆）
10:00〜10:25	トピックマップを用いた人名典拠情報の構築　○研谷 紀夫, 内藤 求（東京大学大学院情報学環, ナレッジ・シナジー）	生物分野における Web API の適用 ○重元 康昌, 桑名 良和, 權 娟大, 菅原 秀明（富士通, 東海ソフトウェア, 国立遺伝学研究所）
10:25〜10:50	国内大学図書館におけるデジタルアーカイブの現状　○鈴木 良徳, 時実 象一（愛知大学文学部）	バイオメタデータベースの構築とその応用 鈴木 智典, ○宮崎 智（東京理科大学薬学部）
10:50〜11:15	博物館における業務情報の共有と IML (Inter-Museum Loan) システムの可能性 ○田良島 哲（国立文化財機構 東京国立博物館）	個々の疾患に特有な関連遺伝子を特定するアルゴリズム ○權 娟大, 菅原 秀明（国立遺伝学研究所 生命情報・DDBJ 研究センター）
11:15〜11:40	自然言語処理を用いた格言引用システムの開発 ○牧野 晃典, 梶川 裕矢（東京大学大学院工学系研究科）	統合システムを有する分散型熱物性データベースの展開 ○山下 雄一郎, 馬場 哲也（産業技術総合研究所 計測標準研究部門）
11:40〜12:05	児童書における文章の長さの変化 ○山本 昭, 日下 陽子（愛知大学文学部）	マテリアルリスク指標データベースの設計と活用　○芳須 弘, 原田 幸明, 藤田 充苗（物質・材料研究機構）
12:05〜13:00	お昼休憩	
13:00〜13:40	論文賞表彰式＋記念講演	—
	セッション A-2 （座長: 阪口 哲男）	セッション B-2 （座長: 石塚 英弘）
13:40〜14:05	テキスト批評の計量化に向けて－書評の計量分析－ ○村井 源, 徃住 彰文（東京工業大学大学院社会理工学研究科価値システム専攻）	論文数・引用数からみたわが国の大学間格差の動向分析 ○根岸 正光（国立情報学研究所／総合研究大学院大学）

時間	A 会場（W935）	B 会場（W934）
14:05〜14:30	村上春樹の初期三部作における構造解析 ○工藤 彰, 村井 源, 徃住 彰文（東京工業大学大学院社会理工学研究科価値システム専攻）	ネットワーク指標を用いた学際的な論文の抽出　○梶川 裕矢, 森 純一郎（東京大学大学院工学系研究科総合研究機構）
14:30〜14:55	文学作品群の特徴的語彙と概念カテゴリーの抽出 — 星新一ショートショートの計量分析— ○佐藤 知恵, 村井 源, 徃住 彰文（東京工業大学大学院社会理工学研究科価値システム専攻）	産学連携に関する Web 情報の分析 －大学サイトの事例研究－ ○柿沼 澄男, 孫 媛, 西澤 正己, 大山 敬三, 根岸 正光（国立情報学研究所）
14:55〜15:05	休憩	
15:05〜15:30	音楽評論論文にみる概念構造の変遷 —ネットワーク中心性を用いた音楽概念の抽出— ○河瀬 彰宏, 村井 源, 徃住 彰文（東京工業大学社会理工学研究科 価値システム専攻）	論文からの解析手順知識の網羅的抽出から分かる解析技術の経年的変化の分析 ○荒木 次郎（(株) 三菱総合研究所 先端科学イノベーショングループ）
15:30〜15:55	心の状態と言語的特徴：ブログにおける商品紹介文の分析 ○斉藤 香里, 村井 源, 徃住 彰文（東京工業大学大学院社会理工学研究科価値システム専攻）	協調フィルタリングを用いた共著関係の予測 ○堀 幸雄, 中山 堯, 今井 慈郎（香川大学総合情報センター, 神奈川大学理学部, 香川大学工学部）
15:55〜16:20	書評中の語を用いた図書に対する感性パラメータ自動設定の最適化 ○原田 隆史, 池内 正明（慶應義塾大学文学部）	世代間共同研究による知識移転のモデル化 ○福田 毅哉, 梅本 勝博（北陸先端科学技術大学院大学知識科学研究科／東京大学物性研究所, 北陸先端科学技術大学院大学知識科学研究科）
16:20〜16:35	休憩	
	セッション A-3 （座長: 梶川 裕矢）	セッション B-3 （座長: 堀 幸雄）
16:35〜17:00	完全合理性と限定合理性 ○安平 哲太郎（産業技術総合研究所研究業務推進部門）	思考展開図を用いた SQL 文の可視化手法 ○村川 猛彦, 中川 優（和歌山大学システム工学部）
17:00〜17:25	情報過程の諸相 ○棚次 奎介（(元) 北九州市立大学基盤教育センター）	サーチエンジン検索結果ページにおける視線情報の分析 ○高久 雅生, 江草 由佳, 寺井 仁, 齋藤 ひとみ, 三輪 眞木子, 神門 典子（物質・材料研究機構, 国立教育政策研究所, 東京電機大学, 愛知教育大学, 放送大学, 国立情報学研究所）
17:40〜19:30	情報交流会（会場: コラボレーションルーム (2F)）	

トピックマップを用いた人名典拠情報の構築
Constructing Authority Control File Using Topic Map

研谷紀夫*1　内藤求*2
Norio TOGIYA,　Motomu NAITOU

*1 情報知識学会 所属　東京大学大学院情報学環
Japan Society of Information and Knowledge, Interfaculty Initiative in Information Studies
the University of Tokyo
E-mail: togiya.norio@iii.u-tokyo.ac.jp,

*2 ナレッジ・シナジー　株式会社
Knowledge Synergy Inc
E-mail: motom@green.ocn.ne.jp

歴史上の人名には様々な「異名同人」などが存在し、歴史的な情報を扱う場合はこれらの統制を行うことが望ましい。現在ではそれらを制御するための共有化されたデータベースは存在しないが、主要な人名に関する構造化されたデータベースを構築し、標準的なデータ記述言語によって構造化された人名典拠情報を構築する必要がある。本研究では、これらの背景を踏まえて、トピックマップ形式を用いて、構造化された様々な事物相互の関係が分かる人名典拠情報を構築した。

Controlling the historical human name is needed ,because there are many instances of a “same person sharing the different name” as a historical figure in historical material. Although no database exists that could be shared to control human name at the current time, there is a need to build structured databases about primary individuals and build sources of name authority information by standardized data description language. With this background, we constructed structured name authority file which user can understand human relationship using Topic Map.

キーワード:人名典拠,文化資源,アーカイブ,トピックマップ Authority Control, Cultural Resource Topic Map.

1.本研究の背景と研究概要

美術館、博物館、図書館などの文化資源保存施設において、各資料に掲載された歴史的人名に関する典拠データベースの構築は重要な課題の一つである。古今東西、古代から現在に至るまで、人の名前は一人の人名に複数存在することが多く、またその表記の方法も異なる事が多い。また、異なる人物でありながら同じ名前を有する事も多く、それらの情報の統制が必要である。特に近年、文化資源をデジタル化して公開するデータベースシステムなどにおいては、検索の精度を向上させるため、これらの人名の統制が特に必要となっている。

また、これらの人名に関する情報はより構造化された標準的な形式で公開することが望ましい。そのことによって、様々な関係性を明確にし、複数のデータベースにおいてその関係性を再現することが可能となる。情報の標準的な構造化は様々な形式が存在するが、本研究ではこれらにトピックマップ（Topic Map）を用いて記述とその表現を行うこととした。本論ではこれらのトピックマップを用いた人名典拠情報の構築方法について解説する。

2.人名・組織典拠情報の基本構造

2.1. プロジェクトの全体像

著者は、主に戦前期の社会的または文化的に重要な人物に関する典拠情報を構築するプロジェクトを行っているが、それらのプロジェクトの遂行において主に三つの課題を設けている。一つは図書館、文書館、博物館などの三館で用いることの可能な典拠情報をデザインし、構築することであり、二点目は、これらの典拠情報の構築にあたって、複数の人間の間でコラボレーションを行う方法論を確立すること、三点目は、これらの情報を標準的な構造化された情報で記述し、アプリケーション間での共有、交換を可能にする。

一点目については拙論においてその基本デザインを記し[1]、二点目についても拙論において、その実験内容を公開している[2]。本論文ではこれらを受けて、三点目の情報の構造化を、トピックマップを用いてどのように実現したかについて述べる。

2.2 明治期の写真と写真師に関する情報

本プロジェクトでは前述したように人名典拠情報を格納する対象を、主に戦前期の社会的文化的に重要な人物を対象とするが、その文化的な事業に関係する人物の典拠情報として、写真師に関する典拠情報を対象としている。

明治期の写真師に関する情報は、これまで様々な人名辞典が出版されているが、資料によって網羅する範囲が異なる点や、同一人物への記述が辞典によって異なることがあげられる。そのため、本プロジェクトでは、写真師に関する様々な参考文献を基本に、改めて各写真師に関する情報を年表方式で書き起こし、それらの情報がどの文献に掲載されているかを確認できる情報を掲載した。

また、写真師は絵師などと同様に撮影に関する技術や手法などが、師弟関係において伝授されることが多い。そのため、写真師を理解する上ではどのような人的なつながりや系統の中で技術を習得したかを理解できる情報を保持することが必要である。そのため、最終的にはそのような系統図の形で情報が構造化、視覚化されることが望ましい。これらの情報の複数人による情報の入力とその独自方式での構造化については前述した拙論について述べている。これらに対して本研究ではこのようにまとめられた情報を、すでに標準化された規格であるトピックマップを用いてどのように構造化と表現を行うことが可能かを明らかにする。

3.人名・組織典拠情報の基本構造とトピックマップへの応用

3.1 トピックマップとは

トピックマップは、ISO/IEC JTC1 SC34 で策定された ISO 標準で、その改訂が現在も続けられているが、2009年4月現在において、その一部は、JIS 規格にもなっている。トピックマップは、情報・知識を分類、体系化し見つけやすくするための技術で、情報空間において本の索引のような役割をする。問題領域における主題、主題間の関係、及び、情報リソースとの関係を、トピック(Topic)、関連(Association)、及び、出現(Occurrence)という構成要素でモデル化してコンピュータ処理可能にする。[3]

トピックマップは、主題を同定/識別するために、PSI(Published Subject Identifier)と呼ばれる仕組みを持っている。PSI は、主題識別子として、主題ごとにユニークな IRI (Internationalized Resource Identifier) を割当てて公開したものである。そのIRIが指し示すアドレス上に置かれ、主題が何を意味しているのか人が理解できるように記述した情報リソースは Published Subject Indicator または、PSD (Published Subject Descriptor) と呼ばれる。本プロジェクトにおいても、PSI を利用することにより、主題が持つ名前や別名で

はなく、IRI によって主題の識別/同定が可能になる。本プロジェクトでは、このようなトピックマップの特徴を人名典拠情報の構築に適応させた。

3.2 作成手順

トピックマップを用いた人名典拠情報の構築は、(1) データ収集、分析、(2) 情報の構造化、(3) トピックマップ作成、(4) アプリケーション開発
の順番で開発を行った。これらの概要について以下より解説を行う。

3.2.1 データ収集、分析

写真史に関する情報は複数の協力者によって収集され、特に分担が必要な情報に関しては、複数の人で必要な情報を集収して行った。これらの情報を複数人で入力し、情報を構造化するアプリケーションについては拙論で述べているが[2]、このように集められた情報をトピックマップ形式で記述することとした。

複数人で集められた情報は**表 1** に表したような、「名前」、「よみ」、「国籍」、「職業」、「生没年」、「略歴」、「出身地」、「居住地」、「所属組織・グループ」、「役職」等の項目である。これらの項目は、トピックマップ作成時のTopics(主題)の候補になる。トピックマップ完成後は、それらの主題に基づいてナビゲートすることが可能になる。

3.2.2 情報の構造化

トピックマップは主に、さまざまな主題 を表現する、トピック(Topics)、トピック間の関係を表現する関連(Associations)、トピックに関連する情報リソースへのリンクである出現(Occurrences)の三要素によって構成されている。

これらの構造を**表 1** の人名典拠情報に次のように適応させた。最初に人名典拠における人名やそれぞれに関係する事項、事項相互の関係性などの全ての事項はトピックマップにおけるトピックと考える。また、人名典拠における、各「人名」と「地名」や、人物間などの関係性は関連(Associations)に適応させる。また、具体的な情報資源へのリンクとなる、「生没年」「略歴」「関係 URL/URI」などの項目が、出現(Occurrences)と対応する。

また、これらのトピック(Topics)、関連(Associations)、出現(Occurrences)にはそれぞれ格納される情報を「型」として類型化することが可能である。最初のトピック(Topics)については、トピックの「型」を表す Topic Types 、関連(Associations)に関しては、関連の種類を定義する Association Types と Association Role Types 、さらに出現(Occurrences)に関しては、出現の種類を示す Occurrence Types などの「型」が設定された。

これらの「型」の具体的な項目としては、Topic Types については、「国」「人」「地名」「組織・グループ」「職業」などのトピックの「型」を示したものをあげることができる。

また、Association Types については、様々な関係を示す「友人関係」「夫婦関係」「師弟関係」「職場関係」「親子関係」を表し、Association Role Types については、関係の生じる各トピックの役割を示し、「上位役割」「兄(姉)役割」「夫役割」「妻役割」「子役割」「弟子役割」などが設定される。

次に、Occurrence Types としては、「よみ(平仮名)名」「よみ(平仮名)苗字 」「ローマ字名前」「ローマ字苗字」「人間関係」「関係期間」「誕生した地」「出典」「役職・役割」「所属」「出典」などが「型」として設定される。

これら各項目に基づいて情報の構造化を行うと、その概要は**図 1** に示された通りである。**図 1** において、四角は、Topic Types を表し、また、線は Association Types を表す。

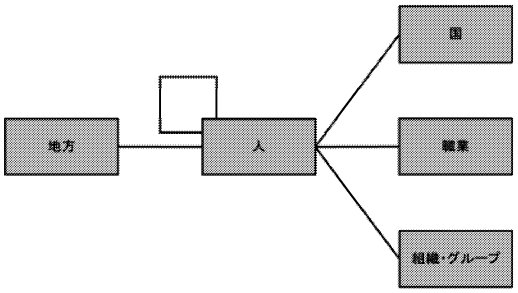


図 1. 人名典拠トピックマップのオントロジ

人型のトピックの間にある親子関係、子弟関係、夫婦関係、友人関係などの関連は、人型のトピックから出て人型のトピックに入る線で表現している。

トピックマップ作成後は、これらの型に基づいた問い合わせ、ナビゲートが可能になり、文字列のみよる検索に比べて意味的な処理が可能となる。

3.2.3 トピックマップの作成

構造化されたデータを元にして、人名典拠トピックマップを生成した。具体的には、OKS (Ontopia Knowledge Suite) [4] の DB2TM 機能を用いてトピックマップを生成した。

DB2TM 機能を使用するために、まず生成するトピックマップの構造を定義する構造化定義ピックマップを作成した。構造化定義トピックマップには、LTM 形式[5]で、トピック、関連、関連役割および出現の各型トピックが記述してある。続いて、CSV 形式のデータをトピックマップに変換する規則を定義するコンフィギュレーションファイルを作成した。最後に、構造化定義トピックマップとコンフィギュレーションファイルを使用して DB2TM のバッチ処理にてトピックマップを生成した。

3.2.4 トピックマップアプリケーションの作成

作成した人名典拠トピックマップを表示するための Web アプリケーションに関しては、OKS の Navigator Framework 機能を利用して開発した。

Navigator Framework 機能は、Java 2 Enterprise Edition (J2EE)、Java サーブレット及び Java Server Pages (JSP) 技術に基づいている。そして、OKS 問合せエンジン、JSP タグライブラリ、Java API から構成されている。開発した Web アプリケーションは、J2EE コンテナのもとで稼働することができ、作成されたトピックマップをトピックや関連に沿ってナビゲートすることができる。OKS Navigator

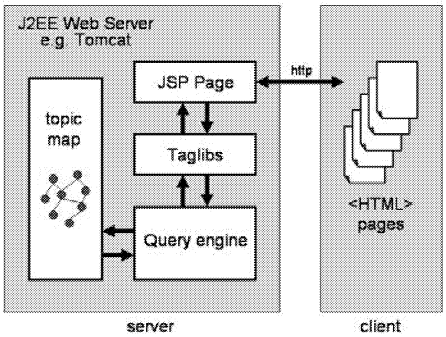


図 2. OKS Navigator Framework のシステム構造

Framework 機能のシステム構造は**図 2** に示した通りである。

3.3 データの表示

作成された WEB アプリケーションは**図 3** に示されているように、作成した人名典拠情報を項目ごとに表示させることが可能である。

トピックの概要は、各トピックに関する、Subject Identifiers、Associations、Internal Occurrences の各情報が表示されている。Associations においては、「子弟関係」「産まれた場所」「活動した場所」などの情報が格納され、表示されており、Internal Occurrences の中には、名前の「よみ」「代数」「関係する人物」「役割」「生年」「没年」「履歴」なおどの情報が掲載されている。

また、これらの中で、Associations に関して

は、**図4**にあるように、グラフ図を用いて、可視化することが可能となった。各トピックはクリックすることによってそれが中心のネットワーク図が表示され、またプロパティを開くことによって、各トピックの詳細な情報を参照することができるようになっている。

4.成果と課題

前セクションで解説したような行程を経て、人名典拠情報をトピックマップで表現し、それらの情報を構造化した上で、それらの関係を表現した。人名典拠における各要素は、概ねトピックマップの基本概念に対応させることが可能となった。

また、人トピックにIRIを割りつけることによって、名前ではなく、IRIによって人の識別が可能になり、マージ機能を利用することで、他のトピックマップのデータと情報を共有することが可能である。また、PSI として公開すれば他のトピックマップやRDFと情報共有が可能となる。ただ IRI として、どのような文字列が適切かは今後の検討事項である。

さらに、トピックマップでの表現で最も課題となったのが、「型」の表現をどのように行うかである。特に、Association Types と Association Role Types に関しては人物相互やそれぞれの事項、要素相互の関係性の他、組織などにおける役割などもこれらのカテゴリーに入る。

人物相互の関係などは予め統制することが可能であるが、役割などについては、それぞれの組織で様々な役職名などが存在する。そのため、これらについて、あらかじめ統制することが難しい。しかし、異なる役割や役職名であっても類似するケースなどがあり、これらの関係性の統制が必要となる。

また、可視化については、**図4**に示されるように、その関係性の可視化することが可能となった。しかし、これらの表示に関しても 3 つ以上の事項のつながりの表示に関しては表示されるトピックの数が多く、認識が難しくなるなどの問題が生じる。これはその他のビュワーについても生じる問題であり、考慮が必要である。

以上が人名典拠情報のトピックマップでの表現の結果であるが、これらのトピックマップ形式を用いることで、文化資源をデジタル化して公開するコンテンツ・システムで実際に利用する際に利便性などの点でどのような問題があるかを検証する必要がある。



図3. 典拠情報の表示

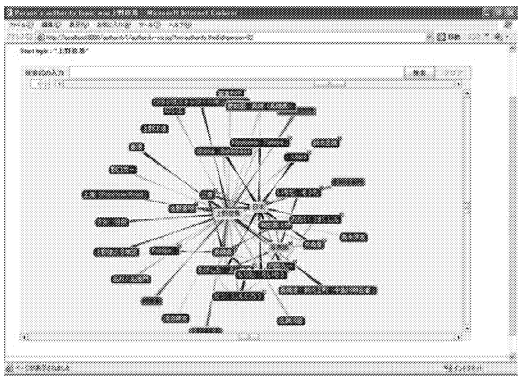


図 4. トピックの可視化

参考文献

[1] 研谷紀夫, デジタルネットワークにおける歴史的人名・組織情報の現状とその利活用モデル, 情報知識学会誌,Vol.18 No.2,2008

[2] 研谷紀夫　川島隆徳２,ネットワーク文化情報資源で活用する人名典拠情報とその共同構築のためのオントロジエディタ

[3] 内藤求、加藤弘之、桐山孝司、小町祐史、瀬戸川教彦、中林啓司、吉田光男、トピックマップ入門、東京電機大学出版局

[4] Ontopia: Software Products: The Ontopia Knowledge Suite, http://www.ontopia.net/solutions/products.html

[5] Ontopia: The Linear Topic Map Notation – Definition and introduction, version 1.3, http://www.ontopia.net/download/ltm.html

表1:人名典拠情報における項目

人典拠情報の項目			TopicMap での対応
名前		漢字(姓/名)	Topic Name
		よみ(姓/名)	内部 Occurrence
		英字(姓/名)	内部 Occurrence
	（複数可能）	世系（複数可能）	内部 Occurrence
		名前の種別（別名雅号など）（複数可能）	内部 Occurrence
国籍（複数可能）			別 Topic を association で結ぶ
性別（複数可能）			別 Topic を association で結ぶ
職業（複数可能）			別 Topic を association で結ぶ
人物 ID			SubjectID
関係 URL/URI		人物 URI	外部 Occurrence
		関係 URL(複数可能)	外部 Occurrence
生没年		生年月日（西暦・新暦のみ）（複数可能）	内部 Occurrence
		没年月日（西暦・新暦のみ）（複数可能）	内部 Occurrence
略歴	略歴　日本語		内部 Occurrence
	略歴　英語		内部 Occurrence
誕生した地　（複数可能）			別 Topic を association で結ぶ
存在した地　（複数可能）			別 Topic を association で結ぶ

管理データ	入力日付　（複数可能）	内部 Occurrence
	最終更新日付、	内部 Occurrence
	種別	内部 Occurrence
	言語コード（複数可能）	内部 Occurrence
	文字コード	内部 Occurrence
	典拠の確立状況	内部 Occurrence
	入力者　（複数可能）	内部 Occurrence

関係（複数可能）	師匠・弟子・知人・父・母・兄・姉・弟・妹・夫・妻・子	Association

国内大学図書館におけるデジタルアーカイブの現状

Digital Archives at University Libraries in Japan

鈴木良徳[1*], 時実象一[1*]
Yoshinori SUZUKI, Soichi TOKIZANE

1* 愛知大学文学部
Faculty of Letters, Aichi University
〒441-8108 愛知県豊橋市町畑町 1-1
E-mail: tokizane@aichi-u.ac.jp

デジタルアーカイブ推進協議会が 1996 年から国内におけるデジタルアーカイブの調査を数回おこなってきたが, 2005 年に同協議会が解散してしまったため, その後調査が行われていない. 現在の状況を把握するため, 2008 年 6 月に国内の大学図書館 716 館を対象に Web によるアンケート調査を実施したところ 222 館（31%）が回答した. それによれば現在デジタルアーカイブを実施している図書館は 57 館で, これは 2004 年調査の 68 館と比べてやや少なかった. しかし Web 調査をおこなったところ回答していなかった 15 館がデジタルアーカイブを実施しており, 合計 72 館が実施していることがわかった.

Japan Digital Archives Association (JDAA) conducted surveys of digital archive activities from 1996 till 2005 when it was dissolved. To learn the current activities, we conducted a survey inheriting prior survey conducted by the Association in 2004. Survey requests were sent to 716 university libraries in June 2008, and received 222 (31%) responses. According to the survey, 57 university libraries had digital archives, which was a little fewer than the result (68) of the JDAA's survey. However, 15 more libraries, which did not respond to our survey, were later found to run archives as well.

キーワード：デジタルアーカイブ, 大学図書館, 電子化, Digital Archives, University Libraries, Digitization

1 本研究の背景

デジタルアーカイブの定義はいろいろ可能であるが, 本研究ではデジタルアーカイブとは,「図書館, 公文書館, 博物館等が収蔵する歴史的価値のある有形・無形の文化財を電子化保存し, Web 上で閲覧可能にするシステム」と定義する. デジタルアーカイブを構築することで,

① 時間, 場所による制約なく資料を公開・閲覧することができる
② 資料を安全に公開することができる
③ 元資料では困難な詳細な調査研究が可能となる

④　データベース化による資料検索が可能となる

といった利点がある.

しかしながら, デジタルアーカイブ構築には様々な問題が存在することも事実である. デジタルアーカイブの構築をさらに発展させるためには, 各機関が抱えている問題を把握し, 解決していかなければならない. このような背景を踏まえ, 著者は全国の大学図書館を対象にデジタルアーカイブに関するアンケート調査をおこなったので以下に報告する.

2 デジタルアーカイブ推進協議会の調査

1996 年に文化庁, 通商産業省 (当時), 自治省 (当時) の支援を得て, 企業 34 社, 10 自治体, 5 団体の会員によりデジタルアーカイブ推進協議会 (JDAA) が設立された[1]. 協議会は, デジタルアーカイブ推進のために, 海外のデジタルアーカイブ状況や諸問題解決のための調査, 広報誌の発刊等の活動をおこなっていた. デジタルアーカイブ推進協議会の歩みについては Web サイト等[1, 2]に記載されている.

同協議会では「デジタルアーカイブ白書」を 2001 年度版, 2003 年度版, 2004 年度版, 2005 年度版の計 4 回発行し, その中で国内のデジタルアーカイブ活動に関するアンケート調査結果を公表している. 同協議会は 2005 年に解散したが, その調査結果が現在でもデジタルアーカイブに関する最も包括的なデータである. その後は同種の調査はおこなわれておらず, 現状が把握できなくなっているので, 著者は最新の状況を把握するためのアンケート調査をおこなうこととした. 同協議会の最後の包括的な調査は 2004 年のもの[3]であったので, その結果と比較するため, 主としてそのアンケート項目を引き継ぎ, これに若干の項目を追加する形でアンケートをおこなうこととした.

3 今回の調査の目的と手法

アンケート調査は 2008 年 6 月におこなった. 調査対象は全国の大学図書館すべてとし, 日本図書館協会発行の「図書館年鑑 2006」[4]から 716 件の図書館を抽出し依頼状を送付した. 大学図書館で分館やキャンパス毎に館が分かれているものについては主要な一館のみに依頼した.

調査はインターネットフォーム「せっかくメールフォーム」[5] を利用し, そこにアクセスし回答を記入していただく方式とした. 依頼状は 06 月 20 日(金)に発送し, 締め切りは 07 月 11 日(金)までとした. ただし, 締切日を過ぎても同年 7 月中に届いた回答は集計対象とした.

4 調査結果

4.1 デジタルアーカイブの実施状況

アンケートの回収数 (222 件), 回収率 (31%)とも協議会調査をわずかに上回った (表 1).

4.2 デジタルアーカイブの実施状況

前述のように, 主たるアンケート項目は協議会の調査項目を引き継いだが, 若

干の項目は今回新規に追加した．それらについては個別に示した．

(1) 実施数

協議会の調査と比較すると，調査時点でデジタルアーカイブを実施している図書館は68館から57館に減少した結果となった（図 1).

実際にアーカイブ数が減少することは考えにくいので，これは回答の漏れや、ばらつきによると思われるが，いずれにせよ顕著な変化はなかったと見られる．ただし回答館には一部の主要大学が含まれていなかったため，別途 Web で調査したところ，さらに 15 館が実施していることがわかった．これらを加えると現在 72 大学でアーカイブが実施されていると考えられる．その一覧は Web に掲載した[6].

なおアーカイブを計画中である図書館は 9 館から 17 館へと増加した．

表 1. アンケート回収数（協議会調査と本論文 2008 年調査の比較）

	協議会 2003 年調査	協議会 2004 年調査	本論文 2008 年調査
アンケート発送数	75	638	716
アンケート回収数	35	184	222
回収率	47%	29%	31%

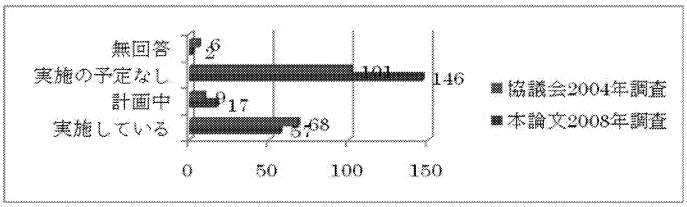


図 1. デジタルアーカイブ実施状況（協議会 2004 年調査と本論文 2008 年調査の比較）

(2) 保存されている資料数

図 2 はデジタルアーカイブで保存している資料数を示しているが，5000 以上のコンテンツを保存している図書館が国立大学で 5 館，私立大学で 2 館あり，逆に 10 以下のコンテンツしか保存していな

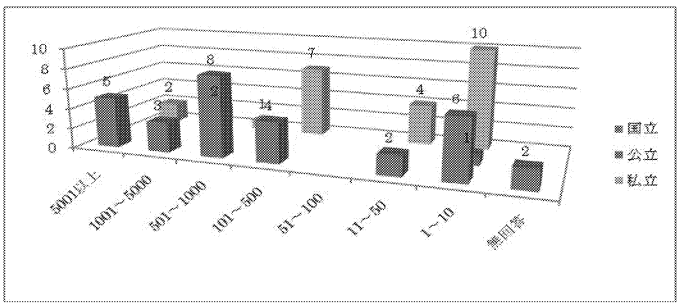


図 2. アーカイブ資料数

い図書館が国立大学で 6 館，私立大学で 10 館という結果となり，各図書館の間に差があることがわかる．ただしコンテンツの数え方は指示していなかったので数え方によってばらつきがあると思われる．

(3) 保存されている資料種類

保存されている資料種類について文書形態の資料と文書以外の形態の資料に分けて質問した（図 3, 4）．文書形態の資料では，貴重本を保存している図書館が 42 館と多く，古文書（手書き資料）が 23 館，古書籍・雑誌・新聞等が 19 館であった．また，文書以外の形態の資料では，絵画・彫刻を保存している図書館が多く，14 館であった．

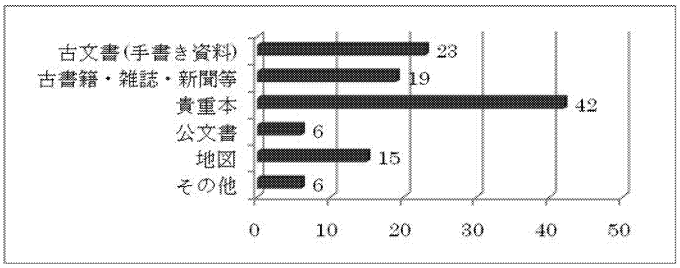


図 3. 資料種別（文書形態の資料）

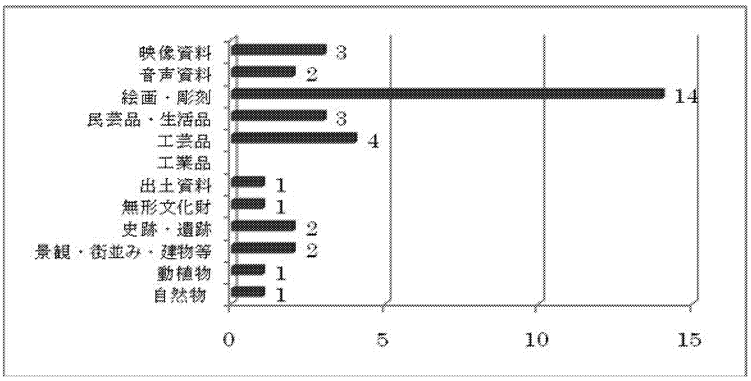


図 4. 資料種別（文書以外の形態の資料）

4. 3 デジタルアーカイブの実施上の諸問題

(1) 問題の種類

図 5 はデジタルアーカイブ導入後の問題についての回答であるが，どの項目も協議会調査より今回調査の方がやや減っているように見える．特に「データのフォーマットの標準化がなされていない」という問題についてはほぼ半分になっており，この点では前進があったものと考えられる．また，今回「長期保存の安定性の不安」という項目を新たに加えたが，これは 23 館が問題として感じていることがわかった．

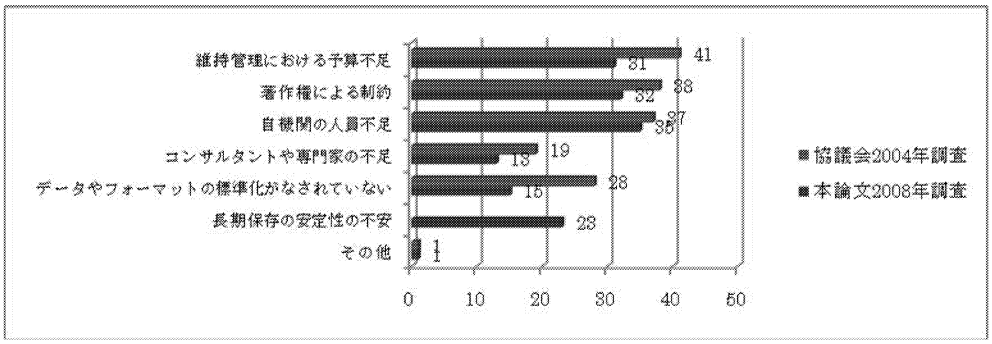


図 5. デジタルアーカイブ導入後の問題（協議会 2004 年調査と本論文 2008 年調査の比較）

(2) 著作権処理（今回新規）

デジタルアーカイブ作成にあたっておこなっている著作権処理について質問をおこなった結果が図 6 である．書面等で権利保有者から許諾を得ている図書館が 66 館と最も多く，次に多いのが，著作権フリーの資料のみを電子化しているというもので 48 館であった．

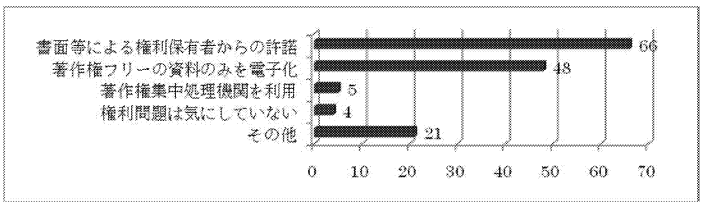


図 6. 著作権問題に対しておこなっていること

(3) 著作権問題についての要望（今回新規）

著作権問題に対して将来的に希望することについての回答は図 7 のとおりである．「書面化された公的マニュアルが必要」と答えた図書館が 104 館と最も多かったほか，「権利問題を扱う専門家が必要」，「著作権法などの整備が必要」などの回答が多かった．

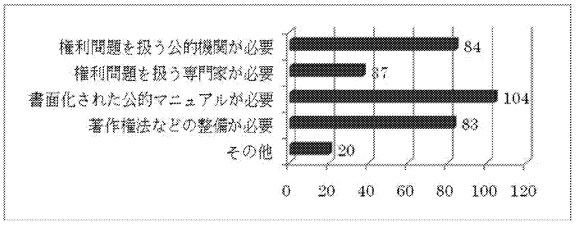


図 7. 著作権問題に関して将来的に希望すること

(4) デジタルアーカイブに関する各種標準の利用状況（今回新規）

デジタルアーカイブにおいて利用している種々の標準においても質問をおこなった(図 8). その結果, ダブリンコアが 20 館と最もよく利用されていることがわかった.

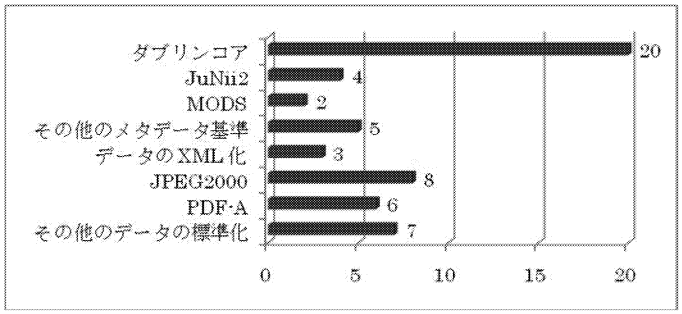


図 8. デジタルアーカイブ標準化について現在利用しているもの

(5) デジタルアーカイブ構築・運営の体制

デジタルアーカイブの構築・運営についての体制について質問をおこなった(図 9). これを見ると, 典型的な図書館は大学において委員会・ワーキンググループを設置し（18 館), 作業は館外の専門的機関に作業を委託する（25 館）か, または学生アルバイトや臨時職員におこなわせている（18 館）という構図になる.

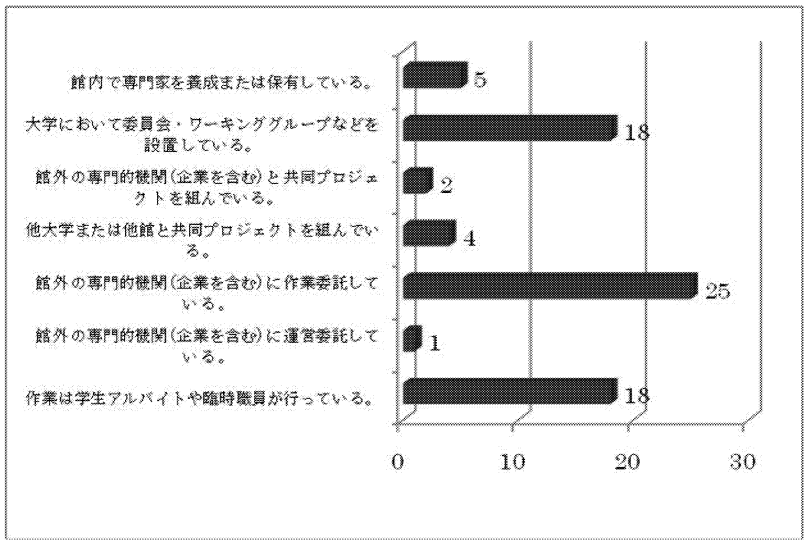


図 9. デジタルアーカイブ構築・運営について現在おこなっていること

5 考察

デジタルアーカイブの新規事業があまり増えず，やや停滞状況と見られることについては今後その理由について調査が必要である．

アーカイブされている文書型資料については貴重本が圧倒的に多く，古文書と古書籍・雑誌・新聞，地図がこれに次いでいたが，これは大学図書館としての特徴と考えられる．また非文書型資料としては絵画・彫刻が多数であったが，これは大学図書館がしばしば大学の博物館・美術館の役割を果たしていることを示していると考えられる．

デジタルアーカイブ事業遂行上の問題点としては，予算・人員は当然のことながら，著作権の制約が大きな問題として認識されていることは協議会の調査結果と変化がない．また今回長期保存の安定性について懸念を持っていることも明らかとなった．

標準化についてはダブリンコアが広く使われていることが明らかになった．しかし JPEG2000 や PDF-A の採用など長期保存を意識した取り組みはまだ少ない．今後個々のデジタルアーカイブのより具体的な実情について掘り下げて研究を進めていきたい．

最後にアンケート調査にご協力頂いた大学図書館の皆様に心より御礼申し上げます．

参考文献

[1] デジタルアーカイブ推進協議会(JDAA)関連情報. http://www.dcaj.org/jdaa/ (閲覧 2009/4/5).

[2] デジタルアーカイブ白書 2005. デジタルアーカイブ推進協議会. 2005.

[3] デジタルアーカイブ白書 2004. デジタルアーカイブ推進協議会. 2004.

[4] 図書館年鑑 2006. 日本図書館協会. 2006.

[5] せっかくメールフォーム. http://fm.sekkaku.net/ (閲覧 2009/4/5).

[6] 日本のデジタルアーカイブ一覧. http://tokizane.jp/tokizane/Shiryo/DigitalArchives.html (閲覧 2009/4/5).

博物館における業務情報の共有と IML(Inter-Museum Loan)システムの可能性

Sharing curatorial infomation among museums and developing “Inter-Museum Loan” system

田良島 哲*
Satoshi TARASHIMA

＊独立行政法人国立文化財機構　東京国立博物館
Tokyo National Museum, National Institute for Cultural Heritage
〒110-8712　東京都台東区上野公園13－9
E-mail: stara@tnm.jp

日本の博物館における調査研究，展示，貸借，閲覧，修理などの学芸業務に関する手順は，個々の館によってさまざまな方式があり，標準化されていない．これは事務の合理化のみならず，外部への情報提供についても多大な不都合を生じている．小論では，博物館における主要な学芸業務である展覧会の「貸借」業務を素材として取り上げて，その実態を概観する．その上で図書館サービスにおけるILL(Inter-Library Loan)との対比でIML(Inter-Museum Loan)という概念を提起し，それが業務を合理化するだけではなく、外部サービスの改善にも役立つことを指摘する．

It is a problem that Japanese museums have no common standards or guidelines for dealing with curatorial works, like registration, loan, or conservation. This report suggests, as an example, possibility of developing Inter-Museum Loan(IML) system inspired by ILL in library service. That will be based on a shared information among museums, and will not only rationalize the process of ordinary job, but also will make effective contribution to public service.

博物館，情報処理,学芸業務，展覧会，ILL
museum, information processing, curatorial works, exhibition,ILL

1 問題の所在

博物館・美術館等（以下「博物館」という）が所蔵，管理する資料に関する情報の処理に関しては，2005年に筆者を含む東京国立博物館（以下「東博」）のプロジェクトチームが「ミュージアム資料情報構造化モデル」を公表して，資料自体が含む情報の構造化と共有を図るための議論の素材とした[1]．

このモデルは単に資料自体から読み取られる情報を構造化するだけではなく，博物館の実務との連関を考慮し，業務を通じて獲得されるデータを再利用可能な情報として取り込んでゆくことに意を払っている．

モデルの編集に当たった村田良二は，「本モデルでは業務支援を視野に入れて，ミュージアムにおいて業務的に発生したデータがそのまま資料の履歴として管理されてゆくことを想定している」と述べている[2]．筆者も同じ立場にあるが，この段階では博物館業務プロセス自体のモデル化はさらに検討が必要であり，後日の課題とした．

その後，東博では，この資料情報構造化モデルに基づいた列品(所蔵品)データベースを構築するとともに，作品・資料の受入，平常展示案の作成など個別の業務を担うサブシステムを開発し，日常の実務を素材に試行を継続してきた．その中で資料情報と業務情報との関連付け，情報項目の整理，用語の共有といった問題点に対して実践的な解決を図ってきた．

小論で提起するのは，実践段階に入った館内業務における情報共有を通じて，そのまた少し先に見えてきた「博物館間における業務情報の共有」という課題であり，業務プロセスのモデル化と博物館間で業務情報を共有することを通じて何が実現できるのか，という問いかけに関する回答の素描である．

2 博物館の業務情報の類型化

2005年の調査によると，日本国内には歴史系，美術系の「博物館」と総称される施設が4470ある[3]．その多くが資料の収集，調査研究，平常展や特別展の開催，カタログや調査報告書の刊行，学術的な要望に対する閲覧などの便宜の供与，所蔵品の画像の提供といったいわゆる学芸業務を行っている．

無論，博物館の規模の大小によって，各業務を細かく分担しているところから，一人の職員がすべての業務を引き受けるところまで仕事のあり方は様々だが，業務の内容に応じた対応の原則と明文あるいは暗黙の手順に基づいて，業務を進めているはずである．

ただ，これら学芸業務の手順を標準化したガイドラインや手引きといった資料は乏しく，それぞれ自館の流儀によって，業務手順や文書を作っているのが日本の博物館の現状である．業務手続きが館同士の間で異なるだけではなく，極端な場合，館内（たとえば担当する資料分野）においてさえ異なる場合も少なくない．本来共有化できるはずの手順の差異に気をつかいながら業務を進めることは，博物館員として本来もっと知的，創造的な仕事に回すべきエネルギーを日々浪費しているのに等しく，幅広い視点から改善が講じられるべき問題である．

現代の博物館の業務情報の標準化を考える上で，必ず参照されるのが，英国Museum Documentation Association(現在は改組してCollections Trust)が策定した博物館のコレクション管理業務手順の標準SPECTRUMである．SPECTRUMは1994年に初版が公開されて以来，改訂を重ねており，現在，Collections Trustのサイトからダウンロードできる SPECTRUM3.1は，400ページ近くにのぼる大部なものである[4].

SPECTRUMは博物館のコレクション管理業務手続きを「受入」「借用」「貸出」など21のカテゴリーに類型化し，それぞれに「手続きの定義」「最小限の標準」「関係する法令」「必要な方針」「具体的な手順」「参考文献」など，個別の館が手続きを定める際に参照するべき情報を項目立てしている．

コレクション管理業務にも国や地域の間で偏差が存在するので，SPECTRUMを引き写しはできないが，このような標準的な業務モデルを参照して，各館の手順を作るこ

とは業務の合理化の上で不可欠である．

3 博物館における「貸借」業務

博物館間での業務情報の共有や流通を考える素材として，重要な業務の一つである「貸借」をとりあげてみよう．

現代の博物館では，特定のテーマによる「特別展」「企画展」を計画し，他の博物館や寺院・神社・個人などの所有者から作品・資料を借用して，展覧会を構築している．特に日本の場合，多くの博物館が収蔵品の不足を，借用を主とする特別展で補っている実情があり，特別展抜きでの博物館運営はむずかしい．

博物館における毎年開催される展覧会の数や，展示される作品・資料の数を推計できる資料は意外に少ない．文部科学省の社会教育基本調査の調査項目にある「博物館における事業実施状況」における「事業」とは「講演会，研究会，学級・講座，映写会等」で，なぜか「展覧会」はないのである．

一つの手がかりとして，展覧会等の情報を継続的に収集・公開している株式会社廣済堂のサイト「Museum Café」の「展覧会・イベント情報」検索のページを利用しよう[5]．年度（4月から3月）を単位として展覧会・イベントの件数を検索してみると，2006年度 2238件，2007年度 2519件がヒットした．すべてがいわゆる展覧会ではないが，毎年1000の単位で展覧会が開催され，当然，展覧会数を大幅に上回る数の作品・資料の貸借が発生していることが想像される．控えめに見積もって一つの展覧会で10件の貸借があったとしても毎年，万の単位で作品・資料の貸借が行われていることになる．

たとえば，東博は国内で最も多くの収蔵品を有する館の一つとして，内外の博物館等へ収蔵品・寄託品の貸出を行っており，2007年度の実績で137施設1302件となっている[6]．東博は例外的に貸出が多い館ではあるが，１館平均10件弱の貸出となる．

博物館資料は原則としてその存在がsocial context- dependantであり，代替性がない．そのため，貸借に際しては，資料の保存状況のチェックとそれに基づく輸送及び展示条件の設定，搬送時における借用館または貸出館の職員の随伴，搬送・展示に際しての損害保険の付与など博物館資料特有の性格に伴なう多くの業務が発生する．

また，業務の進行にあわせて，貸出の依頼書，借用側の施設環境に関する書類（ファシリティ・レポート），貸借両館での内部決裁の書類，貸出の許可書，輸送や保険の契約，貸出時に取り交わす借用証書，貸付簿，貸出品の状態に関する調書など，多岐にわたる文書が作成，授受される．

これらの業務にともなう情報は，再利用や共有が容易であれば，さまざまな方向での活用が想定できるのであるが，今のところ，文書形式や表現方法などが様々な状態で，個別館同士の間で共有されるにとどまっている．また，一旦事務が終了した後の再利用も困難で，類似の業務が発生した場合に手間をかけて情報を再作成しなければならないことが多い．

近年，博物館の運営を取り巻く経済的な情勢が厳しさを増す一方で，博物館からの情報提供に対する期待が高まるという一見相反する現状を考えるならば，個別の館が互いに無関係に苦労を重ねている状況を脱して，業務に関する情報の流通・共有を図ることは急務と言えよう．

4 IMLの可能性

図書館の世界で情報共有が進み，コンピ

ュータ・ネットワークが急速な普及を見た大きな動機のひとつに，ILL(Inter-Library Loan)業務の存在がある．ILLは原資料（書籍等）自体の貸借と複写物など情報の相互融通の二つが大きな柱であるが，電子化以前から行われていた業務の需要を適切に捉えてネットワーク上に展開したところが普及の要因であろう．

ILLの考え方を博物館の間での業務に応用するとInter-Museum Loan(IML)という概念を想定することができる．すなわち「貸借」業務のネットワーク上でのシステム化である．対象は展覧会での作品・資料自体の移動と，情報，特に博物館では画像情報の貸借が主となるだろう．前述したように博物館資料は形態的な多様であり，かつ代替性がない．そのことによって発生する手間・コストがきわめて大きいが，一方でILLの場合，貸借１件あたりのコストは低いが，厖大な数のトランザクションが発生する．したがってコスト×件数として見ると，IMLは実践的課題として，そう突飛なものではない．貸借にともなう連絡と文書作成，資料の授受手続きの標準化と情報の共有は，煩雑な事務の軽減に寄与するはずである．

さらに，貸借業務に関する情報を共有するメリットは，単に事務的な手続きが楽になるという点にとどまらない．貸借にかかっている作品・資料の情報が公開されるならば，それはいつからいつまで，どこで公開されたか（リアルタイムの公開ができれば「公開されるか」）という利用の履歴として蓄積される．逆から見ると舞台となった展覧会と開催館のデータベースとしても利用できる．また，展覧会でカタログが刊行されれば，個別の作品・資料がどの展覧会のカタログに出ているのか，といった情報を網羅的に把握することも可能である．

インターネット普及以来の「博物館におけるネットワークの活用」という課題について，「博物館」から「エンドユーザ」に対する情報の提供という視点が主流で，「博物館の間でのネットワーク活用」という問題が取り上げられることが少なかった．館自身，あるいは複数の館の間でのコレクション情報のコントロールができてこそ，エンドユーザへの情報サービスが充実するはずだというのが正論であるべきだが，そのような視点での議論はまだあまり聞かない．

その意味で今回提示したIMLという発想は，この課題を解決してゆく上での一つの素材となるのではないかと考える次第である．

参考文献

[1] 村田良二：「ミュージアム資料情報構造化モデルの構築」，MUSEUM(東京国立博物館)，No.602，2006.06.

[2] 村田良二：「ミュージアム資料情報構造化モデルによる博物館業務支援と情報共有」，情報処理学会研究報告．DD，[デジタル・ドキュメント]，2006(11)，pp.9-16，2006.01.

[3] 平成17年度文部科学省社会教育基本調査による．

[4] http://www.mda.org.uk/spectrum.htm

[5]http://www.museum-cafe.com/exhibition/index.html （2009 年4 月10日参照）

[6] 平成19年度東京国立博物館自己点検評価報告書http://www.nich.go.jp/hyoka/2007/2007-03-01-01-03.pdf （2009 年4 月10日参照）

自然言語処理を用いた格言引用システムの開発

Development of the Proverb Quoting System using Natural Language Processing

牧野晃典[1*], 梶川裕矢[1],
Kosuke MAKINO, Yuya KAJIKAWA

*1 東京大学
The University of Tokyo
〒113-0033 東京都文京区本郷7-3-1
E-mail: makino@biz-model.t.u-tokyo.ac.jp

本研究は，利用者の書いた任意の文脈に対して適切な引用文を，web上に蓄積された膨大な文書の中から推薦するシステムを開発することを目的としている．今回はこの引用文推薦システム開発の第一段階として，代表的な引用文のひとつである格言に注目した．文脈と格言の間のマッチングを取ることで，文脈に即した格言を推薦するアルゴリズムを複数提案し，性能の比較評価を行った．

The aim of this paper is to develop a system which recommends an attractive quotation for any user's context from huge amount of texts available in the web. As the first step in the development of this Quotation Recommending System, we focus on a major example of the quotation, Proverb. We propose several algorithms recommending a Proverb suitable for a given context by matching context with proverb, and evaluate and compare the performance among them.

キーワード: 自然言語処理，データマイニング， 自動要約， 格言
Natural Language Processing, Data Mining, Automatic Summarization, Proverb

1 序論

1.1 背景

情報技術は知識の蓄積とそこへのアクセスを容易にすることで,我々の情報への接し方を劇的に変化させた．また同時に情報技術は,既に存在する文書群を引用・編集することで,効率的な文書作成を可能にし,知識の生産性を高めた.

この情報の生産性の多寡を決めるのは,新たな情報を生み出す著者自身の能力もさることながら,膨大な文書を入手・理解し，適切な箇所を引用・編集する技術である．これらのうち，情報の入手に関しては検索エンジンを代表として数多くの既存の技術が存在する.また,既存文書の理解を援ける技術としてテキスト自動要約の研究も盛んである.

テキスト自動要約の性能は,如何に重要文を抽出するかに依る．ここで，要約に於ける重要文は従来「その文章の主題を含む」「要約箇所として価値が高い」と定義されてきた．しかし，引用・編集の段になると,その要約された箇所が必ずしも利用者の文脈に,利用者の意図する形でそぐうものであるとは限らない.したがって引用･編集というプロセスを援けるためには,利用者の文脈を考慮した上で既存文書の重要箇所を判断する必要がある．このような，扱う情報の量や質，文脈，ユーザー，社会的背景などの情報変換に伴う環境を考慮する立場は,要約システムに於ける重要な一パラダイムであり，Maria Pintoらの研究[1]ではCommunicational Paradigmとして位置づけられている.

1.2 研究目的

本研究の目的は,利用者の書いた文章に対し､その文脈に対して適切な引用文を，web上に蓄積された膨大な文書の中から推薦するシステムを開発することである.この引用文推薦システム開発の第一段階として,今回は代表的な引用文のひとつである格言に注目した.

引用文推薦システム構築のための足がかりとして,任意の文脈に対して適切な格言を与えるシステムの実装を行う.これを格言引用システムと名付けることにする.

格言引用システムは，たとえば,

「成功を望むのなら，成功談・失敗談数多く読み,思考技術に関する本を読むといいだろう.さらに定期的に散歩する時間をとっておくと良い,思索には欠かせない時間となるはずだ.一人旅も楽しく為になるだろう．」

このような文書に対して,

「世界は一冊の本だ.旅をしないものはその本を一頁しか読めないだろう．」

というアウグスティヌスの格言の引用を推薦するようなシステムである.

これは主に文脈と格言の間のキーワードマッチングによって行う.マッチングの手法は様々であるが，本研究では,各格言の各文脈における内容の合致度を測定する複数いくつかのアルゴリズムを提案し,それらの性能を比較評価することで,より文脈に適した格言を提示できるアルゴリズムを開発し,引用文推薦システムの改良指針とすることを目指す.

2 システム概要

本研究が構築するシステムはひとつの独立した機能を持つシステムである.システムの目的即ちタスクは,与えられた文脈に対して適切な格言をライブラリ中から選択し提示することである.システム全体を図式化すると図１のようになる.

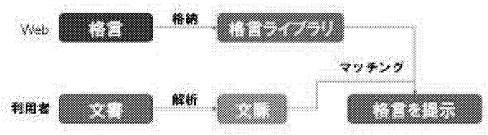


図1 システム図解

3 実装

3.1 仕様要求

Perl5.0及び形態素解析システム「茶筅」を用いて，以下の仕様要求を満たすプログラムを構築した.

－ライブラリの構築・管理
－格言の自動分類
－文脈とライブラリの照合・評価

3.2 アルゴリズム

ライブラリの構築アルゴリズムについては説明を省略する.ここでは格言の自動分類と,文脈とライブラリの照合・評価の手法について,特記すべき事項を記す.

（i）格言の自動分類

格言の自動分類に際しては,従来のカテゴライズ手法に加え,以下の手法を導入した.

－Threshold and Weight Optimization－PSO Classification

Threshold and Weight Optimizationは全ての格言に共通な頻出語（それ,この,である等）を排除するための閾値 Threshold と,先頭語及び同一格言内に複数回出現する語に高い評価を与えるための重み係数 Weight について,その最適な組み合わせをライブラリ更新の際にシステムが自動的に算出し,適用する機能である.また PSO Classification は PSO（Parts Of Speech：品詞）情報を整理し,分類性能が向上するよう重みを配分したものである．（一般名詞,サ変接続名詞,自立動詞の順にスコアが高い,等）

（ii）文脈とライブラリの照合・評価

対象となる文書と格言ライブラリ中の各格言の類似度を以下に示す三通りの手法により評価した.図２に，各マッチング手法の簡単な図解を示す.

（手法１）対象となる文書に出現する語から,最も重要と思われる名詞または動詞を入力文脈のテーマとして単一に決定し，（i）で行った分類をもとに，格言の主題との比較を行った.｛格言の主題とテーマが完全一致する／格言の主題が入力文脈のテーマを含む文字列である／入力文脈のテーマが格言の主題を含む文字列である｝場合によって，それぞれ異なるスコアを与え,これを格言の主題のスコアと乗算した値が,最も高くなるような格言を選択した.

（手法２）対象となる文書を形態素解析し,その文書の有する語ベクトルと格言ライブラリ中の各格言の語ベクトルとの内積を取ることで類似度を計算し、類似度の最も高い格言を選択した.語ベク

トルに含まれる素性は，{単漢字のみ／名詞のみ／名詞及び全ての動詞／名詞及び自立動詞} の四パターンを試行し，今回は，最も精度の高かった「名詞及び自立動詞」を採用した．

（手法３）格言に，その格言を実際に引用している文章を Back-Context データとしてあらかじめ付与しておき，対象となる文書と Back-Context となる文章の類似度を計算する．最も類似度が高い Back-Context データ中において引用されている格言を選択する．類似度の計算方法は手法２と同様である．なお，今回の格言ライブラリ中の複数の格言を引用している文書は Back-Context として採用していない．

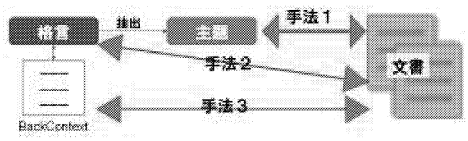


図 2 各マッチング手法図解

3.3 実験

収集した3025本の日本語格言のうち，ランダムに選出した175本の格言を用いて実験を行った．

テストデータにはBlogよりランダムに収集した100本の格言引用文脈を用意し，それぞれの文章に対し、最も内容に沿うと思われる格言を人手で付与し正解データとした．これらの文脈に対してシステムに３．２節（ii）で示した三通りの手法で格言を提示させ，その性能を比較した．

4 結果

4.1 抽出された格言の例

以下に各手法によりテスト文書に対し抽出された格言の具体例を示す．結果の中から，手法によって出力が異なったもの，出力が一致したもの，また特に面白いと思われるものを選出した．

4.1.1 出力の異なった例

入力文書：「突然に舞い降りるチャンス．そんなときは迷わないでください．人の一生はみんな平等です．そのチャンスを掴むのか掴まないのかでその人の人生を左右するのです．」

出力格言（手法１）:「平等はあらゆる善の根源であり，極度の不平等はあらゆる悪の根源である．」

出力格言（手法２）:「いつまでも人を恨んでいてはならない．あなたが恨んでいる人は，人生を楽しんでいる．」

出力格言（手法３）:「機会が二度君の扉を叩くなどと思うな．」

手法３による結果が最も内容的妥当性が高い．手法３により出力された格言を入力文書の後に続けることで機知に富んだ文書の作成が可能となるであろう．

4.1.2 出力の一致した例

入力文書：「其の時の形相を想像するだけで怖いものです．鏡に映った自分の姿の怖さを相手が実感するべきなんでしょうね．叱る，怒ることを否定するものではなく大事なことだと思います．怒る相手にも一分の理，自分と同じ人間の血が流れていることを知らねばなりません．」

出力格言（全手法）:「自分で怒りを抑えるには，他人の怒る姿を静かに観察することである．」

出力格言が入力文書の良い要約となっている。

4.1.3 特に面白いと思われる例

入力文書：「政治家なんてモンは嘘つきばかりで,選挙の前になると減税を連呼して,当選してしまえば舌の根も乾かないうちに増税を推進する.」

出力格言（手法３）:「一つの嘘は嘘である．二つの嘘も嘘である．三つの嘘は政治である.」

出力格言が入力文書に見られる政治批判へのパラドキシカルな回答となっている.

入力文書：「期末テスト．何もしてない．笑えない．最悪の状況に立たされています.」

出力格言（手法３）:「状況？　何が状況だ．俺が状況を作るのだ.」

出力された格言が入力文書の作成者に対する教育的な激励となっている.

4.2 精度比較

評価は次の二通りの方法で行った.

－完全精度:正解との完全一致のみ正解とする.

－拡張精度:予め正解とした格言でなくても，文脈にそぐうものであれば正解とする.

拡張精度を評価に取り入れた理由は,同種の意図を持つ格言が増加すると,正解との完全一致が厳しくなり,結果完全精度はライブラリの充実に反比例した数値を示してしまうためである.

結果は以下の表１に示した通りとなった.ただし，手法１の拡張精度は,出力される結果が支離滅裂のため正解不正解の判定が難しく,評価しなかった.

表１　照合手法によるシステムの精度比較

	手法１	手法２	手法３
完全精度	0.07	0.20	0.34
拡張精度	-	0.48	0.72

表１から明らかなように,Back-Contextを用いた手法３が最も高い精度を示している．加えて,出力された結果を詳しく分析する分析することで、Back-Contextが格言についての解説文の性質を持っている場合。より高い確率で正解となることが分かった.

5 結論

5.1 考察

手法３が高い精度を示した理由は二点考えられる．ひとつは，格言の本文は通常の文脈に比べて遥かに短く語数が少ないため,手法２では類似度が十分に計算できないが,手法３では語数の多いBack-Contextを用いるために類似度の評価が正確になること。もうひとつは,格言によく見られる皮肉表現などの,語の集合のみからは文の真意を判定できない困難を回避できることである.

また格言の自動分類に於ける精度の最高値 0.54 には二通りの解釈の余地がある.

(i)　自動分類は従来の方法の倍近い精度を出すことに成功している.今後ライブラリ･アルゴリズムが充実することでさらに精度を高めることができる.格言の自動分類は実用可能である.

(ii)　格言の分類に特化したアルゴリズムを用いてなお半分程度の精度が限

界である.文脈との照合の前段階から半分の精度では,全体性能の信頼性は極めて低くなる.

後者の解釈は,文脈と格言との照合に格言の分類を利用しないことが可能であるBack-Contextを用いた手法の優位性を示すものであると言える.

5.2 改良

ライブラリのポテンシャルを向上することがもっとも容易かつ効果的な性能改善案である．具体的には,

－登録格言総数を増加する
－語の照合にシソーラスを用いる
－入力文脈の読解精度を向上する
－Back-Contextの量・質を充実させる

などが考えられる．特に，現段階では語の照合はすべて完全一致で行っているため柔軟性に乏しく,精度向上のためにシソーラスの導入は肝要である.

入力文脈の意図をより正しく把握するためには,修辞関係や接続関係を考慮した文節間の連接性や,接続関係に関するヒューリスティクスを取り入れることが有用だろう[2].

語ベクトルの素性についても,今回採用した名詞＋自立動詞に限らず,よりよい選択あるいは組み合わせを探索する余地がある.

また Back-Context の質については,４章の結果から,より解説文に近い性質のものを選ぶべきであると結論する.

5.3 展望

今回本研究で実装した格言引用システムは､それ自体では与えられた文脈に対し適切な格言を返す機能を持つに留まるが，これを拡張した引用文抽出は,他分野に於いても様々な応用の可能性が考えられる．一例を挙げれば，裁判に於いて事例に対して引用すべき過去の判例を自動で選別する判例引用システムなどが考えられる.

また格言引用システム自体も Blog など web に於ける需要は少なくない．本研究で実装したシステムの機能は,格言をランダムで提示する現行の多くのプラグインとは一線を画している.ユーザーの記事に即した動的なサービスとして，展開の余地がある.

参考文献

[1] Maria Pinto and Carmen Galvez: “Paradigms for abstracting systems”, Journal of Information Science, Vol.25, No.5, pp.365- 380, 1999

[2] Samuel W.K. Chan: “Using heterogeneous linguistic knowladge in local coherence identification for information retrieval”, Journal of Information Science, 2000

児童書における文章の長さの変化

Sentence Length in Children's Books

山本昭，日下陽子
Akira YAMAMOTO, Yoko KUSAKA

愛知大学文学部
Faculty of Letters, Aichi University
〒441-8522　愛知県豊橋市町畑町1-1

児童書に短文化傾向が見られるのかを、絵本の15シリーズを対象に調査した。1冊の総文字数を句点など、文章の切れ目の数で割り、1文の平均文字数を出し、年代との相関を調査した。出版年と平均文長の間に負の相関関係が見られるものが多かった。さらに2シリーズについての文長の平均、標準偏差、最長、最短、変異係数を出した。変異係数に時代との相関は見られなかった。

The average sentence length were compared to the publication dates, in Children's books, which had been published through years. The negative correlations between publication date and average sentence length, which signify that the sentences in children's books tend to become shorter. Also, the standard deviation and the coefficient of variation in sentences of a book were studied, in two series. There seemed no tendency in the coefficient of variation.

キーワード: 絵本，文の長さ，児童書，計量言語学，短文化
Children's books, Length of sentence, Quantitative linguistics, Short message

１．はじめに

近年、携帯小説やメールの普及によって、最近の文章は短くなったといわれている。文学作品(1)や大統領の演説(2)等に関する報告がある。これらの原因としては、携帯電話などのメディアの影響、わかりやすさの追求(3)(4)等が指摘されている。本研究では、短文化傾向が児童書でもみられるかを調査した。

２．方法

児童書とは本来、対象年齢は乳幼児から中学生くらいまでであるが、本研究では、幼児～小学校低学年程度のものを対象とした。ベストセラー本や、日本絵本賞や国際アンデルセン賞など、有名な賞を受賞した作品、もしくは、同タイトルで話が異なったり、作者が違う作品を中心に調査する方法も検討したが、対象年齢が異なる図書を扱うと、文章の長さが全く違う為、調査方

法としては適切でない。対象年齢が同一のものを対象にすることも考えられたが、対象年齢が明記されていない作品がある為､断念した｡シリーズ本は、基本的に対象年齢も同じであるし、シリーズ化されるということは、人気があり、多くの人に読まれており、その時代を代表した作品であると考えられる。長く続いている児童書のシリーズ本のうち、絵本に絞り、翻訳か否かを問わず、日本語で発行されているものから､｢うさこちゃん｣､｢ブルーナの本｣、｢くまのアーネストおじさん｣、｢おさるのジョージ｣、｢くまくん｣、｢１４ひき、１１ぴきのねこ｣、｢ぐりとぐら｣、｢ノンタン｣、｢ばばばあちゃんのおはなし｣、｢かこさとしおはなしのほん｣、｢名作アニメ絵本｣、｢ゆうたくんちのいばりいぬ｣､｢カルちゃんとエルくん｣、｢おばけえほん｣ の１５シリーズ本を調査対象とした。

３．結果

それぞれの１冊ごとの本文全体の総文字数を数え、｢句点｣、｢！｣、｢括弧｣など、文章の切れ目の数で割り、１文の長さの平均を出した。以下、本研究では､｢１文｣を、句点、！、括弧等の文の区切りの数で、１冊の総文字数を割ったものを指すこととする。なお、漢字の部分は、ひらがなに直したときの読み方の数で数え、数字は、今回の調査では、２桁までしか出てこなかった為、全て１文字として数えた。

【例】自由（じゆう）→３文字
２ひき→３文字
１４ひき→３文字

図 1-16 では、シリーズごとに、横軸に発行年、縦軸に平均文長を示す。相関係数は小数点第２位を四捨五入した。

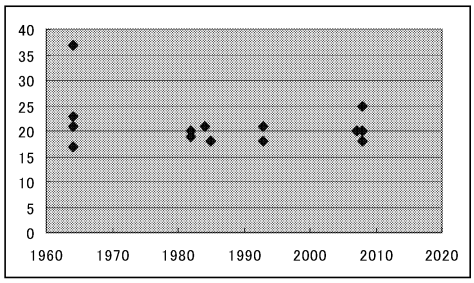


図１　うさこちゃんシリーズ
相関係数 -0.5 出版年間 45 年(1964~2008)

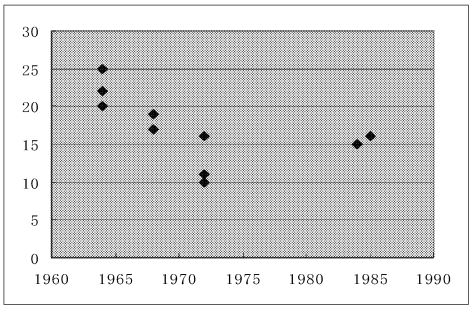


図２　ブルーナの本シリーズ
相関係数 -0.5 出版年間 22 年(1964~1985)

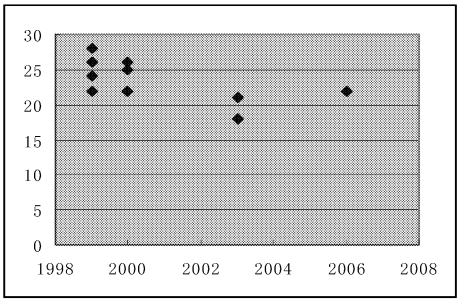


図３　おさるのジョージシリーズ
相関係数 -0.7 出版年間 8 年(1999~2006)

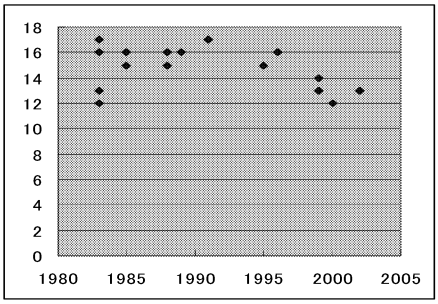


図４　くまのアーネストおじさんシリーズ
相関係数 −0.4 出版年間 20 年(1983~2002)

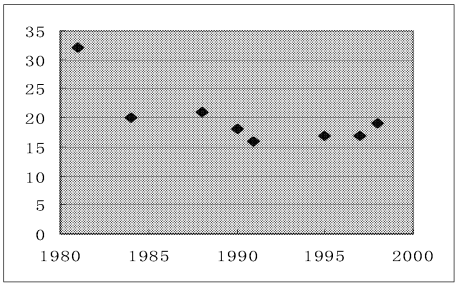


図７　ばばばあちゃんのおはなしシリーズ
相関係数 −0.7 出版年間 18 年(1981~1998)

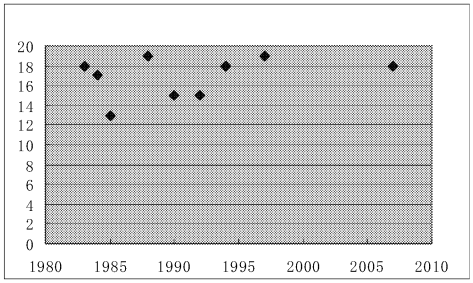


図５　１４ひきシリーズ
相関係数 −0.3 出版年間 25 年(1983~2007)

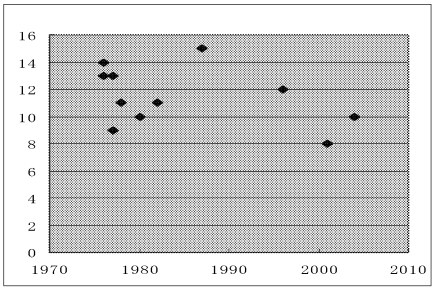


図８　ノンタンシリーズ
相関係数 −0.4 出版年間 29 年(1976~2004)

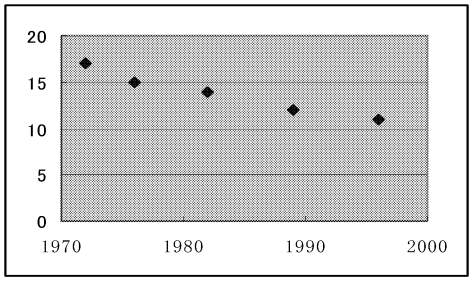


図６　１１ぴきのねこシリーズ
相関係数 −1.0 出版年間 25 年(1972~1996)

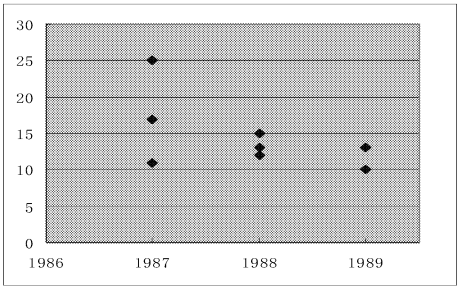


図９　ノンタン（赤ちゃん版）
相関係数 −0.6 出版年間 3 年(1987~1989)

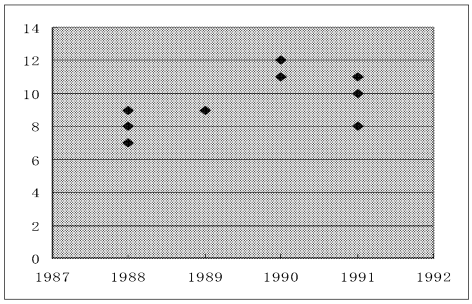


図 10　ゆうたくんちのいばりいぬシリーズ
相関係数 0.6 出版年間 4 年(1988~1991)

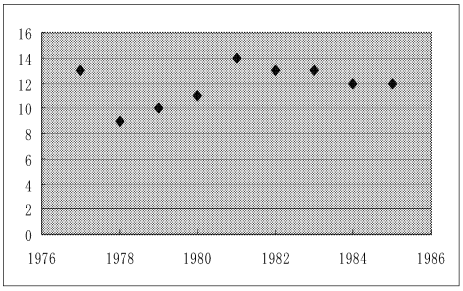


図 13　くまくんの絵本シリーズ
相関係数 0.4 出版年間 9 年(1977~1985)

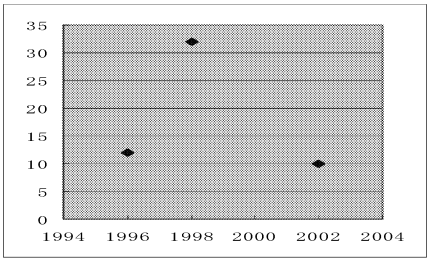


図 11　カルちゃんエルくん
相関係数 −0.3 出版年間 7 年(1996~2002)

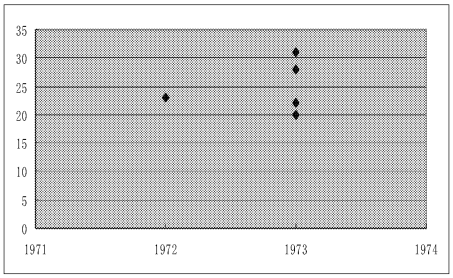


図 14　かこさとしおはなしのほんシリーズ
相関係数 −0.5 出版年間 2 年(1972~1973)

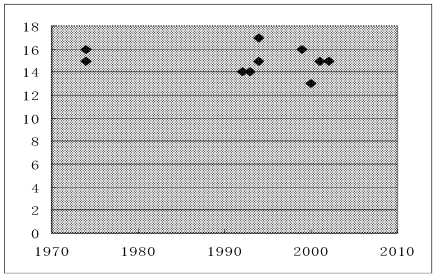


図 12　おばけえほん
相関係数 −0.2 出版年間 29 年(1974~2002)

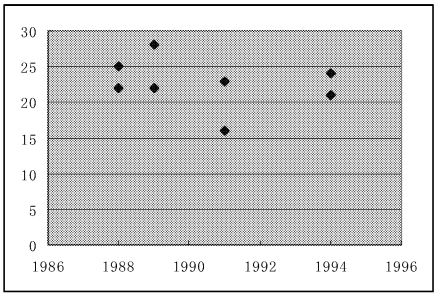


図 15　名作アニメ絵本シリーズ
相関係数 −0.3 出版年間 7 年(1988~1994)

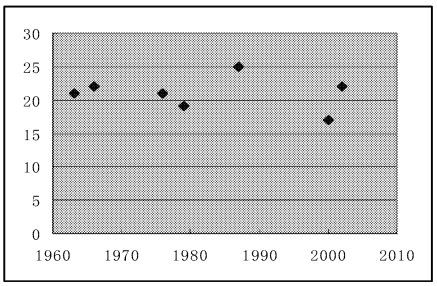


図 16　ぐりとぐらシリーズ
相関係数 −0.2 出版年間 40 年(1963~2002)

次に、１文の長さのばらつきに関して調査した。手作業ですべてを調査するのは難しいので､今回調査した中で、もっとも長いシリーズ本である「うさこちゃん」シリーズと、相関関係の値であまり文の長さに変化が無かった「くまのアーネストおじさん」シリーズのいくつかの文の長さの平均、標準偏差､最長､最短､変異係数を出した。

「句点」､「！」､「括弧」などで区切れる１つずつの文章の文字数を数えた。（以下、本調査では、その１つずつの文の文字数を「１文章」と呼ぶ。）文の長さの平均、標準偏差、最長、最短、変異係数を求めた。平均とは、総文字数を「句点」､「！」､「括弧」の数で割ったものでり､本研究における､「１文の長さ」である。これらを表１－２に示す。

表１「うさこちゃんシリーズ」の文長

年代	1964	1964	1964	1964	1982	1993	2007	2008	2008	2008	2008
平均	16	16.7	26.3	18.4	16.5	19.7	19.1	16	18.9	17.8	23.3
標準偏差	6.4	8.1	11.6	10.9	10.4	10.6	8	8.1	7.9	8.8	7.6
最長	30	49	52	58	44	48	35	30	37	35	39
最短	4	4	12	6	2	6	6	4	7	5	10
変異係数	0.4	0.5	0.4	0.6	0.6	0.5	0.4	0.5	0.4	0.5	0.3

（平均、標準偏差、変異係数は、小数点第２位を四捨五入。）

表２「くまのアーネストおじさんシリーズ」の文長

年代	1983	1985	1988	1999	2002
平均	12.9	14.6	16.1	13.3	12.7
標準偏差	7.9	7.2	9.3	5.8	7.3
最長	36	33	45	32	41
最短	2	3	2	3	3
変異係数	0.6	0.5	0.6	0.4	0.6

（平均、標準偏差、変異係数は、小数点第２位を四捨五入。）

６．考察

文長に関しては、負の相関係数の値が出たシリーズは７、正の相関係数の値が出たシリーズは４、相関係数の絶対値が 0.4 以下のものが５であった。

外国のシリーズ絵本は、すべて、負の相関係数の値が得られた。

カタカナの文字に注目して、調べてみると、明治や大正の本などでは、ひらがなで書かれている言葉でも、昭和・平成などに書かれた比較的新しい本では､カタカナで書くことによって、文の長さが短くなったり、長くなったりと、変化することがあるが、今回調べた児童書では、出版年が最高で 45 年しか空いていなかった為、そのような変化はみられなかった。

児童書は“よいしょ よいしょ”や“ゴシ ゴシ”,“バシャ バシャ”などの繰り返し言葉が目立ち、ノンタンシリーズでは主人公の“ノンタン”､おさるのジョージシリーズでは、主人公の“ジョージ”など、小説のような“私”という一人称ではなく、固有名詞が頻繁に出てくるなど、文長に影響を与える要因が考えられる。

今回の調査の代表的な児童書として、２シリーズの文の長さの平均、標準偏差、最長、最短、変異係数を出した。変異係数においては２シリーズとも、違いは見られなかった。

７．結論

児童書において。１文の短文化傾向が認められた。時代変化による作品中での１文章の長短の差はあまりなかった。

参考文献

[1]星川法子.形態素解析による若い作家の小説の特徴の研究.園田学園女子大学 情報コミュニケーション学科卒業論文 13p. (2005)
http://www.sonoda-u.ac.jp/dic/kenkyu/2005/53.pdf（参照 2008-12-03）
[2] 笹井常三.英文ライティング･ハンドブック.研究社,1989,p.92-93.
[3]山本恒.学校教育は情報社会と真剣に向き合ってほしい. i-Net vol.22, p.2-5 (2008)
http://www.chart.co.jp/subject/joho/inet/inet22/inet22-1.pdf （参照 2008-12-03）
[4]杉山晴信.“米国のビジネス通信文における短文化傾向の特徴―センテンスの長さと構造に関する文体統計学的考察―”. 独協大学英語研究 vol.47,p.41-63.(1997)

生物分野におけるWeb APIの適用

Web API for Biology

重元　康昌[1*], 桑名　良和[2], 權　娟大[3], 菅原　秀明[4]
Yasumasa SHIGEMOTO, Yoshikazu KUWANA, Yeondae KWON, Hideaki SUGAWARA

*1 富士通株式会社
Fujitsu Limited
〒144-0035　東京都大田区南蒲田2-16-1 NOFテクノポートカマタセンタービル
E-mail: yshigemo@genes.nig.ac.jp
2 株式会社東海ソフトウェア
Tokai Software
〒410-0011　静岡県沼津市岡宮 1463-86
E-mail: ykuwana@genes.nig.ac.jp
3 国立遺伝学研究所
National Institute of Genetics
〒411-8540　静岡県三島市谷田 1111　生命情報・DDBJ 研究センター
E-mail: yekwon@lab.nig.ac.jp
4 国立遺伝学研究所
National Institute of Genetics
〒411-8540　静岡県三島市谷田 1111　生命情報・DDBJ 研究センター
E-mail: hsugawar@genes.nig.ac.jp

生物の研究現場ではさまざまな解析プログラムやデータベース、文献などの情報があり、それらを組み合わせて使うことが多い。そこで、国立遺伝学研究所では 120 以上の機能について Web API を開発し公開した。これによって、利用者は国立遺伝学研究所の大規模システムが手元にあるかのようにさまざまなサービスを自由に組み合わせて利用することができるようになるとともに、組み合わせのノウハウをワークフローとして共有することも可能になった。本稿ではこの生物分野における Web API の適用事例を紹介する。

There are various analytical programs, databases and bibliography in biology and they are often integrated and used. Therefore, we at National Institute of Genetics (NIG) developed Web APIs of over 120 functions and opened them to the public. As a result, users of NIG systems can combine several services as if the user has the large scale system. Furthermore, the know-how of the combination can be shared as a workflow. We introduce the application of Web API in biology.

キーワード: Web サービス, 相互運用性, ワークフロー, バイオインフォマティクス, ナビゲーション
Web service, interoperability, workflow, bioinformatics, workflow navigation

1 はじめに

生物学の研究者はある１つの研究・解析を行うためにさまざまなシステムを利用する。このときに最も時間がかかり、大変な作業の１つは、あるシステムの出力を別のシステムの入力とするために変換することである[1]。この変換は手作業によるコピー・ペーストで実施されることが多く作業が煩雑であるとともに、手順が残らないためノウハウが共有されない問題もある。そこで、これらの問題を解決するために国立遺伝学研究所 DNA Data Bank of Japan (DDBJ) では生物分野における Web APIサーバ (WABI)を公開した。また、複数のWeb APIを部品としてつなげたワークフローやある解析の結果から次に遷移可能なサービスを列挙するナビゲーションシステムも構築した。

2 生物分野におけるWeb API

2.1 Web APIとは

APIは Application Programming Interface の略で、プログラミング言語から利用可能な機能や関数の集合を表す。そして、Web APIはAPIの実体がインターネット上にあるAPIであり、それらのAPIを組み合わせることで新たなサービスを作るマッシュアップ技術も注目されている。

2.2 生物分野におけるWeb API

生物分野ではさまざまなデータベースや解析プログラム、文献などがあり、これらを統合して使うことが多い。そこで、DDBJでは、プログラミング言語から利用可能なSimple Object Access Protocol (SOAP[2]) や REpresentational State Transfer (REST) サーバを開発し公開した。このサーバでは、表1に示すようなデータベース検索・取得や解析など21サービスを公開している。

表1 国立遺伝学研究所で公開しているWeb APIサービス

サービス名(Web API数)	サービス概要
DDBJ(7), ARSA(4), GetEntry(44)	20以上の生物データベースにおけるキーワード検索やデータ取得
Blast(6), ClustalW(4), Fasta(5), VecScreen(4)	相同性検索や多重アライメントなどのDNA配列やアミノ酸配列を用いた解析
Gib(11), Gtop(3), GTPS(8), GIBV(8), GIBEnv(1), GIBIS(1), SPS(2)	DDBJオリジナルデータベース(微生物／ウィルスゲノム、環境由来配列、ゲノムのORF再評価、タンパク構造)
TxSearch(5), RefSeq(1), GO(3), Ensembl(4), OMIM(2), NCBIGenomeAnnotation(4)	DDBJ以外の研究所で開発されたデータベース

2.3 Web APIの利用方法

国立遺伝学研究所で公開しているWeb APIの1つであるRESTサービスの利用方法を説明する。RESTサービスは PerlやJavaなどのプログラミング言語で利用することが可能である。

国立遺伝学研究所で構築しているDNAのデータベースDDBJには1億件を越すDNA配列が収録され、各エントリにはユニークIDであるアクセッション番号が割り振られている。Perl言語を用いてRESTサーバからアクセッション番号を指定して1件のDNA配列のエントリを取得する例を図1に紹介する。

```
use LWP::UserAgent;
$ua = new LWP::UserAgent;

# 接続情報の設定
$req = new HTTP::Request POST => 'http://xml.nig.ac.jp/rest/Invoke';
$req->content_type('application/x-www-form-urlencoded');

# パラメータの設定 (アクセッション番号の指定)
$req->content('service=GetEntry&method=getDDBJEntry&accession=AB000100');

# RESTサーバへ接続
$res = $ua->request($req);

# 結果の表示
print $res->content;
```

図1 RESTサービスの利用例

2.2.4 SOAPとREST

国立遺伝学研究所では、当初SOAPサービスを開発し公開していたが、全DNA配列であるゲノム情報やそれらの情報を用いた解析結果の情報が大きくなり、SOAPでは大規模なデータを扱う際にパフォーマンスの劣化が激しいことからRESTサービスも追加して公開した。例えば大腸菌ゲノムのエントリは約13MBあり、SOAPで取得する場合は19.4秒かかるがRESTで取得する場合、5.1秒しかかからない。さらに、RESTではさまざまな言語で同じように利用することができるとともにWebブラウザやtelnet, wget コマンドを使っても利用することができる。一方、SOAPも利用者のコーディング量が少なくてすみ、拡張性に優れているなどの利点があるため、両方継続して公開している。

3 ワークフロー

3.1. ワークフローとは

ワークフローは一連の手続きの流れのことである。国立遺伝学研究所で公開しているWeb APIを部品として組み合わせることでワークフローを簡単につくることができる。国立遺伝学研究所では以下に示す8つのワークフローを公開している。

・ホモロジーワークフロー
DNA配列を用いたヒトと他の生物種の比較・解析

・ヒトの染色体の遺伝子ワークフロー
染色体ごとの遺伝子数をカウントしダウン症などの疾患との関連探索
・DNA多型ワークフロー
人種により異なるDNA多型解析
・BLAST-ClustalWワークフロー
配列解析プログラムBLASTとClustalWを組み合わせたワークフロー
・スプライシングワークフロー
スプライシング（遺伝子の構造）の観点でヒトとマウスの遺伝子の比較解析
・OMIMワークフロー
ヒトの疾患関連遺伝子と他の生物の遺伝子との比較
・SNPワークフロー
ヒトの遺伝子とSNP(個人差がある部位)の関係抽出

3.2 ホモロジーワークフロー

Web APIを組み合わせたワークフローについて説明するために、例としてホモロジーワークフローを取り上げる。

3.2.1 ワークフロー概要

DDBJにはヒトをはじめチンパンジーやマウスなどさまざまな生物種のDNAデータが登録されている[3]。このデータを利用することで複数の生物種における遺伝子の比較解析を行うことができる。そこで、ヒトの遺伝子と同様な遺伝子を持つ生物種とその類似度をひとまとめに検索することが可能なワークフローを構築した。図2にUMLを用いたワークフローのアクティビティ図を示し、表2にワークフローに ABO を入力した場合の結果例を示す。なお、ABO は血液型を決定する遺伝子名である。

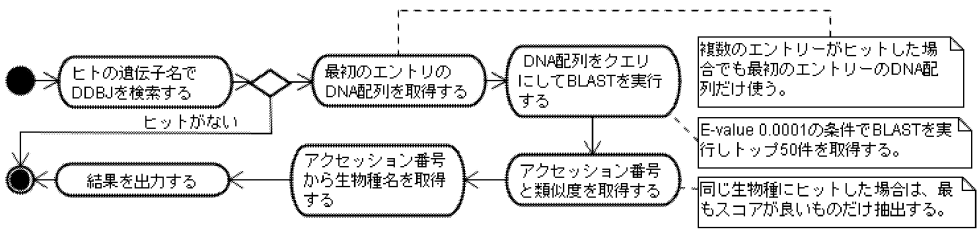


図2 DNA配列解析ワークフロー

表2 ワークフローの結果例（ヒトの遺伝子ABOの結果例）

生物種名	DDBJ ID (アクセッション番号)	DNA配列の類似度
ゴリラ	AY138476	98.69
カニクイザル	AF100981	95.79
アカゲザル	AF071830	95.66
チンパンジー	AF071831	97.90
アヌビスヒヒ	AF019417	96.37

3.2.2 ワークフローの手順

ワークフローの手順を説明する。まず、ARSAシステム[4]を用いて入力されたヒトの遺伝子名でDDBJデータベースを検索

する。次にその検索結果から最初のエントリのDNA配列をgetentryシステムによって取得する。そして、そのDNA配列を元に他の生物種のDNA配列と類似しているものを検索するBLASTシステムを実行し、その結果から生物種名の取得などを行い整理することで表2のような結果を取得することができる。ここで使われているARSAやgetentry、BLASTはWeb APIである。

4 ワークフローナビゲーションシステム

4.1 ワークフローナビゲーションシステムとは

Web APIを組み合わせることでホモロジーワークフローのようなさまざまなワークフローを作成することができる。しかし、Web APIは120以上もあるため、利用者がそれらをどのように組み合わせてよいか分からないこともある。また、さまざまな部品(Web API)を試行錯誤して組み合わせながらワークフローを構築するような利用方法も想定される。そこで、このような利用形態に対応するために、Web APIをWebブラウザで実行することができ、さらにその実行結果を用いて次に遷移可能な部品を列挙するワークフローナビゲーションシステムを開発した。

4.2 ワークフローナビゲーションシステムの構成

ワークフローナビゲーションシステムは、以下の3つの部品で構成される。

・メタ情報
各Web APIのカテゴリ情報やメソッド名、パラメータ名などを定義する。

・辞書
各Web APIの入出力情報やカテゴリ情報を標準化するための辞書。入出力情報を辞書で共有化することによって、あるWeb APIの出力の型と合致する入力情報を受け付けるWeb APIを検索することによって、次に遷移可能なWeb APIを列挙することが可能になる。

・Webインタフェース生成システム
メタ情報や辞書を参照することによってWeb APIをWebブラウザ上で実行することが可能なページを自動的に生成する。

4.3 ワークフローナビゲーションシステムの利用例

ワークフローナビゲーションシステムで複数のWeb APIを遷移する例として、1件のDDBJエントリを取得し、BLASTを用いて解析する手順を図3に説明する。図3-1はDDBJデータベースからDDBJアクセッション番号を指定して1エントリを取得するページの例である。この画面はメタ情報や辞書を参照して自動的に生成される。適当なアクセッション番号を入力し、ボタンをクリックすると、図3-2に示すように実行結果とその実行結果を用いて次に遷移可能なサービスが列挙される。そして、その列挙されたサービスのいずれかをクリックすると図3-3に示すように再びその実行画面が自動生成されるとともに、前の実行結果から抽出されて画面に入力される。

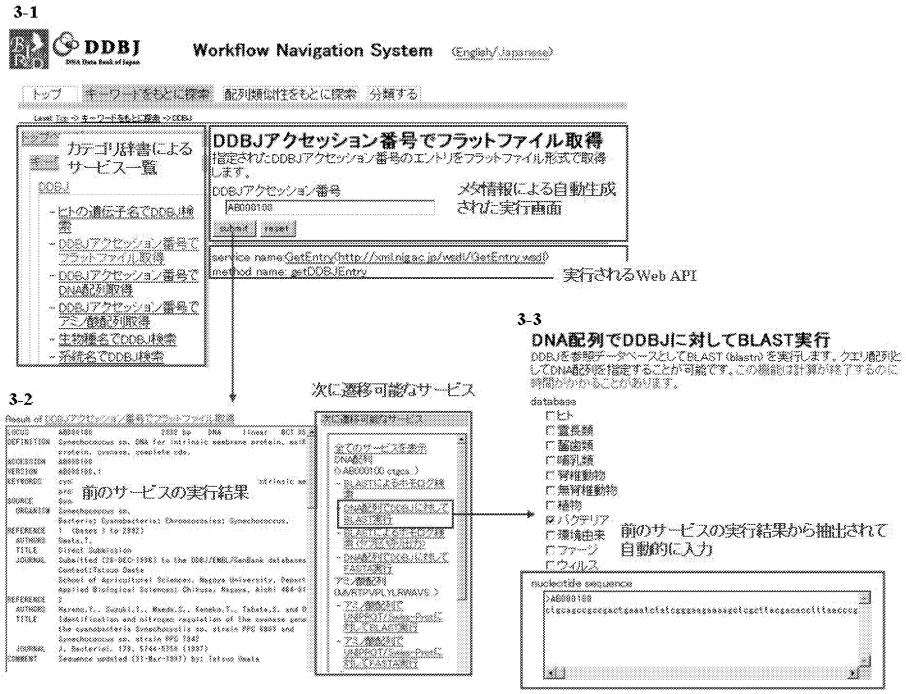


図3 ワークフローナビゲーションシステム

5 おわりに

ワークフローナビゲーションシステムを利用することにより、Web APIをただ単にプログラムから利用するだけでなく新たなワークフローを提案することもできるようになった。より複雑なワークフローにも対応できるように、今後もWeb APIを拡充し、ワークフローナビゲーションシステムの機能を拡張していきたい。

謝辞

本研究開発は、科学技術振興機構(JST) BIRD 事業の支援で実施しています。

参考文献

[1] Stein,L. (2002) Creating a bioinformatics nation. *Nature*, 417, 119-120.

[2] Sugawara,H. and Miyazaki,S.(2003) Biological SOAP servers and Web services provided by the public sequence data bank, *Nucleic Acids Res.*, 31, 3836-3739.

[3] Miyazaki,S., Sugawara,H., Ikeo,K., Gojobori,T., and Tateno,Y.(2004) DDBJ in the stream of various biological data, Nucleic Acids Res., 32, D32-34.

[4] Sugawara,H., Ogawara,O., Okubo,K., Gojobori,T., and Tateno,Y. (2008) DDBJ with new system and face, Nucleic Acids Res., 36, D22-D24.

バイオメタデータベースの構築とその応用

Development of Bio-metadatabase and its applications

鈴木　智典[1], 宮崎　智[1]
Tomonori SUZUKI, Satoru MIYAZAKI

*1 東京理科大学
Tokyo University of Science
〒278-8510　千葉県野田市山崎2641
E-mail: tomoszk@rs.noda.tus.ac.jp, smiyazak@rs.noda.tus.ac.jp

我々は，現在インターネット上で公開されている生物学的データベースのURLや公開形式，検索ページのCGIパラメータなどを網羅的に集めたデータベースのデータベース(バイオメタデータベース)を構築し，データ検索から初め，遠隔操作によって複数のサイトの解析ツールを組み合わせて研究に応じた解析ワークベンチの動的な構築を試みている．メタデータベースの概要とワークベンチの事例について報告する．

To realize efficient utilization of data resources on the Internet, we developed the database for databases named “Bio-metadatabase”. Bio-metadatabase is not only the collection of URL for publicly available databases but also the collection of their access methods including parameters accepted by CGIs provided by the sites. We also expect that Bio-metadatabase will be a core resource to design the semantic Web services and allow us to create user-oriented work-benches for bioinformatics data analysis. In this paper, we described various aspects of our Bio-metadatabase and some ideas with this database such as the one-stop query data retrieval system from multiple sites and the workbench for evolutional analysis of influenza viruses.

キーワード: メタデータ，バイオインフォマティクス, CGI
Keywords：metadata, bioinformatics, CGI

1 背景

各種生物でゲノム単位の研究が行われるようになり，生物学の世界でもさまざまな情報の電子化が促進され，インターネットを通して共有されるようになった．また，従来の生化学実験を主体とした手法のみならず，先の多大なデータに基づくコンピュータ内シミュレーションによる生物学の普及も著しい．ゲノム科学に代表されるこうした新しい生物学においては，インターネットに散在しているデ

ータの活用が欠かせない．データを効率よく収集するためには，コンピュータに不慣れなあるいは非専門家であっても比較的容易にデータ収集ができる仕組みを構築していくことが急務であろう．そのためには，インターネット上で公開されているデータベース[1][2][3][4]について，レコードの書式や検索方法についての情報を統一的かつ網羅的に蓄積したデータベースのデータベース（メタデータベース）の構築と活用が重要となる．しかし，生物学を中心とするデータについては，google[5]やYahoo[6]に代表されるロボット検索によるリンク集様のものがあるものの，実際の検索手法や書式，そのサイトの有用性や信頼性まで含めてまとめている事例は非常に少ない．そこで，我々は，Nucleic Acid Research（NAR）[7]のデータベース特集号に掲載されたデータベース公開サイトを中心に，実際の公開セット名や検索手法を網羅的に蓄積したデータベース（バイオメタデータベース）[8]の研究開発を行っている．特に，検索手法情報の1つとしてCGI [9][10]への引数等を蓄積しており，この情報を用いていくつかのページを動的に連鎖させた新しいワークベンチインタフェースの試作も行っている．次節以降では，バイオメタデータベース構築のパイプラインを紹介し，運用状況や利用法を記述する．

2 データベースの構築

2.1 構築手法と現状

メタデータは，人手による分析と各サイトがインターネットに公開している検索ページのHyper Text Markup Language（HTML）のプログラムによる分析をもとに獲得し，フリーのデータベース管理システムMySQL [11]で構築したデータベースに蓄積している．このバイオメタデータベースにおいては，キーワード検索，相同性検索，登録番号（ID）による検索などの検索ごとに対象となっているデータセットをメタデータベースの1レコードとして生成した．したがって，バイオメタデータベースのレコード数は，NARに収録されているデータベース数の3倍以上となっている．また，各サイトでは，もともとは1つのデータベースであったものを，学名や遺伝子などの観点から分割した上で別名をつけて検索対象のデータセットとして提供していることから，これまでのメタデータベース中の名称とNARデータベース特集号や学術論文上のデータベース名とを対応づけることが困難であった．そこで，バイオメタデータベースに登録した2万7千件余りのデータベースを見直して，文献で使われているデータベース名をもとに1,128件に分類した．この分類ごとに，データベースの概要，総レコード数，利用者からみた使用感について，専用の研究補助者によるオリジナルサイトの利用を通した調査を行い登録している．以下にレコードの作成と登録のためのGUI（査定システムの一画面）を示しておく（図1）．

図2は，「新規登録」の登録用画面である．バイオメタデータベースの応用として，幾つかのページを連鎖させるためには，各々のページの出力形式についての知見が必要である．本GUIを通して，出力形式の分類と登録を行っている．出力形式については，この画面とは別に，分類用のインタフェースを設けている．

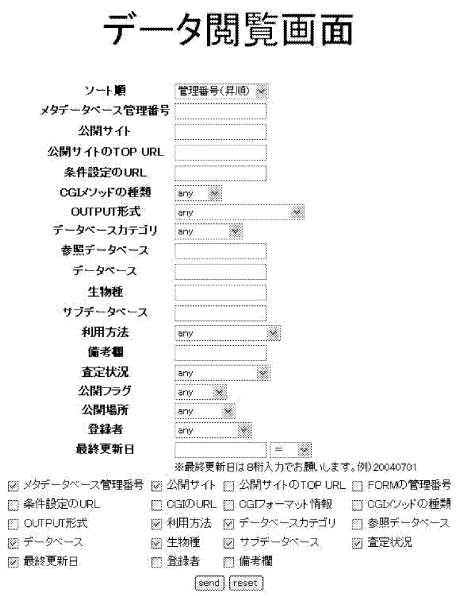


図1 管理者のための登録用レコード更新のためのGUI. 各種条件を設定して登録したレコードを検索することができる.



図2　新規レコード登録用の画面. 多くの項目は選択形式で入力することができる. また, CGIフォーマットは自動解析をした後, 登録者が査定している.

また, 一般利用者がオリジナルサイトについてのコメントを記述できるような仕組みも用意している.

2.2 利用者用インターフェス

2.2.1 インタラクティブアクセス

本システムの公開系には, 主に2つのアクセス方法を実装した. 1つ目はこの節で説明するようなインタラクティブなものである. インタラクティブアクセスでは, さらに, キーワード検索とリスト分類型の2つの用意している. 図3に公開用のホームページの概要をまとめた.

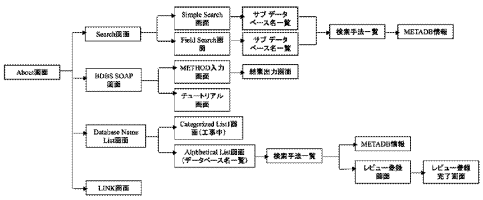


図3 利用者インタフェースの概要

また, 図4および図5にキーワード検索条件設定の画面と, メタデータベースの1レコードの表示例をそれぞれ示す. 利用者が指定したキーワードは, 日本語で記載したデータベースの概要やPubMed [12]のアブストラクトなど, 日本語・英語交じりのテキストを検索するため, 日本語キーワードは英単語へ, 英語の場合は日本語への自動翻訳を行い, 検索をしている.

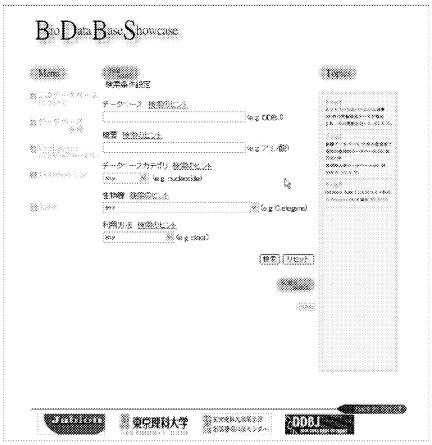


図4　検索条件の入力画面の１つ



図5　検索結果レコードの表示例

2.2.2 ウェッブサービス

本データベースは，データベースのリンク集に留まらず，各データベース用の検索画面のCGIの情報そのものも蓄積している．それらを利用して複数の公開サイトを連結させて，目的に応じたワークベンチを構築できる可能性がある．一方で，こうしたワークベンチは，インタラクティブの利用に留まらず，プログラムから複数のキーワードによる多種の検索をおこない，大量のデータを一括処理するようなツールとしての期待もあろう．そこで，プログラムからの利用を支援する目的で，メタデータベース自身をSOAPサービスで公開している．SOAPアクセスを利用する際の関数やパラメータについてはチュートリアルにて学べるようにしている（図6）．

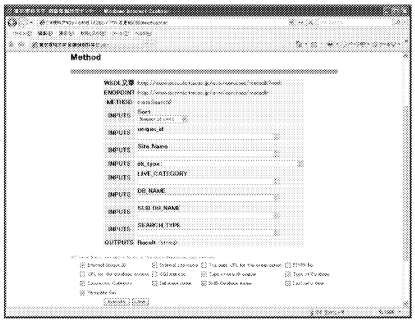


図6　SOAPサービスチュートリアルの一例

3 運用状況と課題

図7の数字は，2009年2月末の本データベースの状況である．

条件	件数
MET ADB総レコード数(削除されたデータも含む)	47,834
MET ADB総レコード数(削除されたデータを含まない)	29,118
今年度更新されたレコード	9,964

公開フラグ	公開場所	件数
公開	一般	21,065
	Selected	4,757
	微生物	584
非公開	一般	2,692
	Selected	0
	微生物	0

※上記の件数は削除されたレコードを含まない件数

図7　蓄積レコード数一覧

総計で47,000以上のレコード登録をしているが，公開しているのは，21,000件となっている．図中で，「selected」としているのは，データベースそのものがレフリー付きの学術雑誌に掲載されており他の学術論文での引用数が比較的おおいもの200データベースについてのレコードを示している．同様に，「微生物」は微生物情報に関連したデータを対象としているデータベースに関するレコード数である．ホームページが更新されると同時に，公開されているデータベースへのアクセスページも更新される可能性が高い．これらの更新状況を機械的に把握する仕組みとして我々は，「パトロール機能」を研究開発してきた．ページ内のHTMLタグの種類や総数を基準にそのページが更新されたどうかを判別し，管理者に知らせる機能である．図8にこの機能によって把握できたページの変遷状況をしめす．この結果をみると，おおよそ1/3のページが1年以内に何らかの更新をしていることが示唆できる．

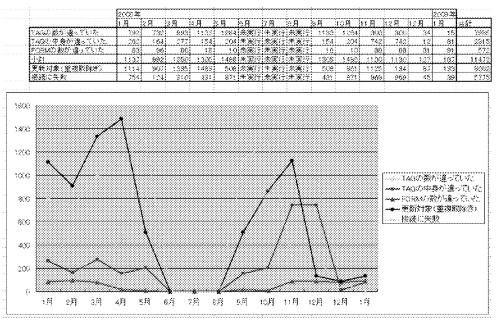


図8　パトロール機能の運用状況

インターネットの変遷に応じてレコードを迅速に更新する新しい仕組みが必要となっている．パトロール機能により変更箇所の目安が付けられるものの，ページ内の変更箇所を機械的に特定することはむずかしく，マニュアル操作が必要となるためである．もう一つの大きな課題は，メタデータベースの検索結果の評価に関連したものである．冒頭で述べたように，本システムは，専門家ではない研究者への検索サービスを構築することを1つの目的としている．そのため，検索結果に対応して，いわば，推奨度のようなものを付加し，利用者が閲覧する順序を支援することが求められてきている．現在，そのデータベースを公開しているページの被リンク数やPubMedアブストラクトでの引用回数などを数量化して評価することを考案中である．また，メタデータベースを利用した研究用ワークベンチの事例として，DNA結合タンパク質[13]の研究を支援するための網羅的なデータベース検索や，インフルエンザゲノム情報[14][15]の統合利用のためのデータベース検索を開発中である．

謝辞

本研究は，平成18年度に，科学技術振興機構バイオインフォマティクス推進事業「バイオ基幹情報資源の高準化と共用化」の分担研究として採択され支援頂いた．

参考文献

[1] National Center for Biotechnology Information (NCBI)：http://www.ncbi.nlm.nih.gov/

[2] European Bioinformatics Institute (EBI)：http://www.ebi.ac.uk/

[3] DNA Data Bank of Japan (DDBJ)：http://www.ddbj.nig.ac.jp/

[4] Protein Data Bank Japan (PDBj)：http://www.pdbj.org/index_j.html

[5] Google：http://www.google.co.jp/

[6] Yahoo：http://www.yahoo.co.jp/

[7] Nucleic Acids Research (NAR)：http://nar.oxfordjournals.org/

[8] Bio Data Base Showcase (BDBS)：http://www.ps.noda.tus.ac.jp/biometadb/index.html

[9] Wall, L.; Christiansen, T.; Orwant, J.（近藤嘉雪訳）：「プログラミングPerl」，オーム社，760 p.，2002.

[10] Gundavaram, S.; Birznieks, G.（田辺茂也，大川佳織訳）：「CGIプログラミング」，オーム社，496 p.，2001.

[11] MySQL：http://www-jp.mysql.com/

[12] PubMed：http://www.ncbi.nlm.nih.gov/pubmed/

[13] 田村隆明：「新 転写制御のメカニズム」，羊土社，283 p.，2002.

[14] Fleischmann,R.D.; Adams,M.D.; White,O. *et. al.*：“Whole-genome random sequencing and assembly of Haemophilus influenzae Rd”, Science, pp. 496-512, 1995.

[15] Genome Information Broker (GIB)：http://gib.genes.nig.ac.jp/

個々の疾患に特有な関連遺伝子を特定するアルゴリズム

Identification Algorithm of Associated Genes Specific to a Particular Disease

權 娟大[1*], 菅原 秀明[1]
Yeondae KWON, Hideaki SUGAWARA

[*1] 国立遺伝学研究所 生命情報・DDBJ 研究センター
Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics
〒 411-8540 静岡県三島市谷田 1111
E-mail: yekwon@lab.nig.ac.jp, hsugawar@genes.nig.ac.jp

MEDLINE アブストラクトのような生物医学文献から疾患に関連する遺伝子を探索する手法が多数提案されている．疾患と遺伝子の関連の度合いは，その疾患名と遺伝子名が共出現する文献の数で測定することができる．しかし，この方法では，他の複数の疾患と関連のある遺伝子が上位に順位付けされる可能性があるので，特定の疾患に対して特有的に関連のある遺伝子を識別することはできない．本稿では，文献集合から特定の疾患に関連する遺伝子群を抽出し，それらを特有性の観点から順位付けするアルゴリズムを提案する．これにより副作用の可能性が小さい関連遺伝子を識別することが可能になり，効率的な創薬に貢献できる．提案手法は遺伝子間の用語共起頻度に基づいて疾患との推移的な関連性についても考慮する．更に，公開されているデータセットを用いて提案手法の精度評価を行う．

There are a lot of interests on extraction methods for finding associations between diseases and genes from literatures such as MEDLINE abstracts. The strength of association between a gene and a disease can be measured by the number of articles in which the gene and the disease co-occur. However, this method cannot identify specific genes to a particular disease because a highly ranked gene may have association with other diseases. In this paper, we propose an algorithm that extracts a group of associated genes with a given disease and prioritizes them in terms of their specificities to the disease. This enables the identification of associated genes that are expected to have fewer side effects, which contributes to efficient drug developments. Our proposed method incorporates transitive associations between the disease and genes based on the frequency of co-occurrence of gene terms. Furthermore, we evaluate the precision of ranking algorithm using a public dataset.

キーワード: 特有性, 関連遺伝子, ランキングアルゴリズム，用語共出現, tf-idf
specificity, associated gene, ranking algorithm, term co-occurrence, tf-idf

1　はじめに

近年，MEDLINE などの文献情報から生物学的知識を抽出する研究が多く行われている [1, 5]．創薬支援を目的とした情報抽出では，特定の疾患に対して特有的に関連する遺伝子を文献中の記述を用いて絞り込むことができれば，効率的な新薬開発につながることが期待される．例えば，アルツハイマー型認知症に対して有効な薬を開発する場合，過去の文献を解析することによってアルツハイマー型認知症に特有的に関連がある遺伝子として APP（amyloid beta (A4) precursor protein）が抽出されれば，APP を対象としてウェット実験を行うことで新薬が開発できる可能性が高いと考えられる．ある疾患に特有な遺伝子とは，その疾患と深く関連があり，他の疾患とは関連性をもたないような遺伝子のことである．これは副作用をできるだけもたない遺伝子を抽出するために必要な概念である．

そこで，本研究では，与えられた疾患に特有的に関連する遺伝子群を生物学文献データから抽出し，それらを特有性に応じて順位付けするアルゴリズムを開発することを目的とする．提案手法は，文献中での用語間の共起関係に基づき疾患と遺伝子の間の関連性の抽出を行う．また，遺伝子間の用語共起頻度を用いて遺伝子のネットワークを構築することにより，疾患と直接の共起出現をもたないような遺伝子との推移的な関連性を考慮する．これらの情報から，各遺伝子について関連する疾患の個数を表す指標値を求め，この値の昇順により順位付けを行う．

MEDLINE から疾患に関連する遺伝子を探索する手法はこれまでに多数提案されている．文献 [2, 4] では，遺伝子，疾患，たんぱく質，化合物などの生物学用語間の関連性を抽出するシステムを開発している．文献 [2] は用語の組合せの出現頻度を利用し，文に対する規則ベースのスコアを加えることにより，用語間の関連性を評価している．文献 [4] は共起頻度による相互情報量に基づき関連性を評価している．文献 [5] は，用語ベクタの表現法，関連遺伝子の順位付けアルゴリズム，利用可能な語彙について様々な選択を同一のベンチマーク上で比較評価し，その結果として，idf (inverted document frequency)，1-SVM，eVOC，MeSH 語彙の有効性を主張している．文献 [3] では，共起出現に基づく遺伝子間のネットワークを構築し，その中からハブとなる遺伝子を関連遺伝子として抽出する手法を提案している．

上記以外にも多くの関連性評価指標が提案されているが，これらはいずれも特定の疾患との関連性の強さに重点を置いており，副作用の可能性の大小については評価値に反映されない場合がある．例として，相互情報量で疾患 d と遺伝子 g の関連性を計算する場合を考える．$p(d)$，$p(g)$ をそれぞれ d，g の出現確率とし，$p(d,g)$ を d と g の共起出現頻度とする．文献数が 10，$p(d) = p(g) = 5/10$，$p(d,g) = 1/10$ のとき，d と g の相互情報量は次式で得られる．

$$\log \frac{p(d,g)}{p(d)p(g)} = \log \frac{1/10}{5/10 \cdot 5/10}$$

しかし，この指標では，g が d と共起しない残りの 4 つの文献で 1 つの同じ疾患と共起する場合とそれぞれ別の 4 つの疾患と共起する場合を区別できない．本研究では，前者の方が d に特有的であるとみなし，これらを区別する指標を用いて上位に順位付けすることを目的とする．

以下，2 章で提案手法について述べる．次に，3 章で提案手法を用いてある疾患の関連遺伝子を抽出する実験を行い，本手法

の有効性を述べる．最後に，4 章でまとめと今後の課題を述べる．

2　提案手法

本研究では，与えられた疾患と特有的に関連をもつ遺伝子を効率的に抽出することを目的とする．このため，副作用を引き起こす可能性がより少ない遺伝子を上位に順位付けする．例えば，ある疾患 d に対して，遺伝子 g_1 は d と高い関連をもつが，他の多くの疾患とも関連をもつとする．一方，遺伝子 g_2 は d との関連性は弱いが，他の疾患とは関連をもたないとする．このとき，g_2 を g_1 より上位に順位付けすることがある．これは，d に対する創薬開発において g_2 の方がより良いターゲットであると考えられるためである．

2.1　用語間の関連性の抽出

文献データベースとして PubMed アブストラクト，遺伝子名辞書として NCBI Gene，疾患名辞書として MeSH（Medical Subject Headings）を使用する．まず，これらの用語辞書を利用して，PubMed アブストラクトから各遺伝子，各疾患の出現表をそれぞれ作成する．次に，出現位置の情報を利用して二つの表を結合することにより，遺伝子-遺伝子と遺伝子-疾患の共出現表を作成する．

我々は，ある文書中での二つの用語の共出現がそれらの間の関連性を示すと考える．二つの用語の共出現の範囲は，同一文書，同一段落，同一文など様々な選択が考えられる．一般に，広い範囲を取る程，再現率（recall）は高く，適合率（precision）は低くなる．本稿では，副作用の可能性を過剰に評価しないように適合率を優先することとし，同一文中での共出現を採用した．更に，自然言語処理技術を用いてより精度の高い関連用語を抽出することが可能であるが，計算時間を要するために大規模データには適用が難しい．

次に，同義語の問題について，遺伝子の同じ実体に対して複数のシノニムが存在するが，これらは同一のものとして扱われるべきである．本研究では，用語出現表作成後に用語辞書を用いて遺伝子シンボルを公式シンボルに正規化した．

2.2　特有性の計算

特定の疾患に対する各遺伝子の特有性を tf-idf（term frequency-inverted document frequency）手法と同様の考え方を用いて定義する．

まず，tf 項について述べる．この項では特定の疾患と遺伝子の共出現の頻度を評価する．d と関連のある遺伝子の集合を $g_1, \ldots, g_n$ とする．疾患 d と遺伝子 g_i が共出現する文書の数を $n_d(g_i)$ としたとき，g_i の d に対する tf 項を次式で定義する．

$$tf_d(g_i) = \frac{n_d(g_i)}{\sum_{i=1}^{n} n_d(g_i)}$$

創薬の文脈においては，tf 値が高いことは g_i をターゲットにすることが効果的であることを意味する．

次に，idf 項について述べる．この項では特定の疾患と遺伝子の共出現の特有性を評価する．つまり，d 以外に他の多くの疾患と共出現する遺伝子は，d のみと共出現する遺伝子よりも低く評価する．これは創薬の文脈では，関連をもつ疾患の数が少ない遺伝子の方が副作用の可能性が低いことを表している．遺伝子 g_i と共出現する異

表 1: アルツハイマー型認知症の関連遺伝子

遺伝子名	tf	idf	sp
APOE	0.229	2.91	0.667
APP	0.145	3.28	0.476
PSEN1	0.028	5.63	0.158
BACE1	0.028	4.70	0.131
CEL	0.165	0.74	0.121

なる疾患の数を $ad(g_i)$，すべての疾患の数を m とする．このとき，g_i の d に対する idf 項を次式で定義する．

$$idf_d(g_i) = \log \frac{m}{ad(g_i)}$$

d に対する g_i の特有度 $sp_d(g_i)$ は次式で求められる．

$$sp_d(g_i) = tf_d(g_i) \cdot idf_d(g_i)$$

特有性の計算例として，アルツハイマー型認知症（Alzheimer Disease）の関連遺伝子を表 1 に示す．CEL は Alzheimer Disease と多く共出現するため tf 値は高いが，他の多くの疾患と共出現するため idf 値が低い．一方，PSEN1 は Alzheimer Disease と共出現する頻度は CEL よりも小さいが，共出現する他の疾患の数が少ないため idf 値が高く，全体として PSEN1 の方が高い sp 値を示している．

2.3　遺伝子間の推移的な関連性

本節では，「文書中である疾患 d と共出現しない遺伝子 g' が他の遺伝子 g と頻繁に共出現し，かつ g が d と共出現するのであれば，g' と d は関連をもつ」という仮説に基づき，特有度に推移的な関連性を取り入れる．

まず，遺伝子間の関係の強さを共出現の頻度を用いて定量化する．遺伝子の集合を $g_1, \ldots, g_n$ とする．遺伝子間の共出現表から $n \times n$ 行列 A を作成する．A の i 行，j 列の要素 $A_{i,j}$ を次式で定義する．

$$A_{i,j} = \frac{oc(g_i, g_j)}{\sum_{k=1}^{n} oc(g_i, g_k)}$$

$oc(g_i, g_j)$ は g_i と g_j が共出現した文書数を示す．但し，$oc(g_i, g_i)$ は g_i が出現した文書数とする．

疾患 d と g_i が共出現をもち，更に g_i と g_j が共出現をもつ場合を考える．このとき，g_i は自身が関連する疾患に加え，g_j が関連をもつ疾患とも何らかの関連をもつとみなす．このような推移的な関連性を考慮した g_i の特有度 $tsp_d(g_i)$ を次式で定義する．

$$idf'_d(g_i) = \log \frac{m}{\sum_{k=1}^{n} A_{i,k} \cdot ad(g_k)}$$

$$tsp_d(g_i) = tf_d(g_i) \cdot idf'_d(g_i)$$

上式において $k = i$ のときが直接的な共出現による特有度，すなわち，$sp_d(g_i)$ に相当し，それ以外のときが遺伝子間の間接的な共出現関係による特有度を表す．ここで，異なる k に対して，同じ疾患が複数の $ad(g_k)$ に含まれることがあるが，それらは別々の要因によって d と関連をもつと考える．

$tsp_d(g)$ の値が大きければ，g が d と高い関連性をもち，他の疾患とあまり関連をもたないことを意味する．従って，$tsp_d(g)$ の値が大きい程 d に対する特有性が高く，g をターゲットとして新薬開発を行っても副作用の可能性が低いことが分かる．関連遺伝子抽出アルゴリズムは，d が与えられたとき，各 g に対して $tsp_d(g)$ の値を求め，この値の降順で遺伝子の一覧を出力する．

遺伝子間の関連性は推移的に定義されるように拡張できる．すなわち，g_i と g_j が関連をもち，g_j と g_k が関連をもつならば，

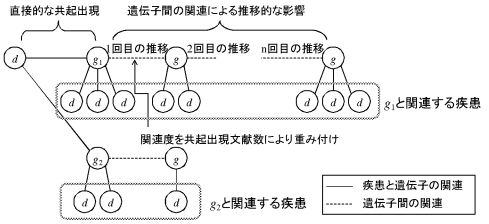


図 1: 遺伝子間の推移的な関連性

g_i と g_k は関連をもつ．このとき，g_i と g_k の関連の強さを $A_{i,j} \cdot A_{j,k}$ で計算する．従って，$A^n = A^{n-1} \cdot A$ の計算を繰り返し行うことで，任意の遺伝子間について n 回の推移的な関連を考慮した関連の強さを得ることができる（図 1）．n 回までの推移的な関連を考慮した特有度は，$idf'_{d,g_i}(g_k)$ の計算式において $A_{i,k}$ の代わりに $\sum_{m=1}^{n} A_{i,k}^m$ を使用することで求められる．

3　実験

3.1　評価方法

遺伝子辞書には，NCBI Gene のヒトのデータ（ftp://ftp.ncbi.nlm.gov/gene/DATA/GENE_INFO/Mammalia/）を使用して，シノニムシンボルを含む 104,165 エントリを作成した．疾患辞書には，NLM MeSH（http://www.nlm.nih.gov/mesh/filelist.html）を利用して，24,767 エントリを作成した．文献データベースには，MEDLINE/PubMed の 2003 年から 2007 年までの 2,871,769 件を使用した．

評価用正解データは，CTD（Comparative Toxicogenomics Database, http://ctd.mdibl.org/）を利用した．CTD は文献から人手または半自動的に生成した遺伝子と疾患の関連データを提供している．

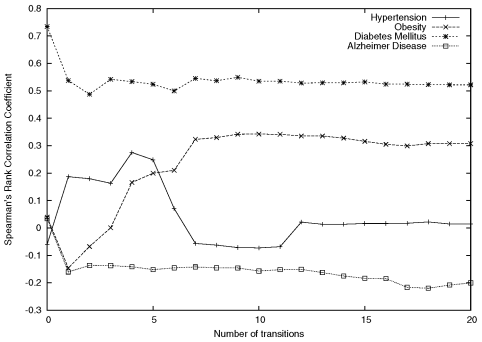


図 2: 評価結果

Alzheimer Disease, Diabetes Mellitus，Hypertension, Obesity の 4 の疾患に対して，CTD をもとに関連遺伝子を作成した．作成方法は，ある疾患の各関連遺伝子に対して，その疾患以外に関連する疾患の個数を抽出し，この値の昇順で順位付けを行った．この正解データは CTD の関連データから 2.2 節で述べた idf 項による特有度を求めたものである．

提案手法の有効性を評価するために，sp の値に基づく順位と n 回（$n = 0, \ldots, 20$）までの推移的関連を考慮した tsp の値に基づく順位を正解データと比較する．正解データとの順位の一致度は，それぞれの上位 30 件の遺伝子リストの Spearman 順位相関係数を求めることによって評価する．sp の結果を基準とすることで，推移関係の繰り返し回数によってどの程度精度が向上したかを評価できる．

3.2　結果

それぞれの疾患に対する評価結果を図 2 に示す．

順位相関係数の値は全体的に低いが，これは提案手法により機械的に生成した関連

表 2: 肥満病の関連遺伝子

CTD	$tsp(n=0)$	$tsp(n=1)$	$tsp(n=10)$
AGRP	IMPACT	IMPACT	IMPACT
PLIN	MC4R	MC4R	MC4R
NPY2R	LEPR	LEPR	LEPR
GHRL	UCP1	UCP3	UCP3
MC3R	UCP3	ADRB3	AGRP
PIGS	MCHR1	MC3R	MC3R
MC4R	MC3R	AGRP	MCHR1
MCHR1	AGRP	ADIPOR2	ADRB3

データには多くの擬陽性を含むことによる．図 2 から，遺伝子-遺伝子間の推移的な関連による特有性への影響を考慮することで，推移回数が少ない範囲で結果が変動していることが分かる．

肥満病に対する CTD から作成した関連遺伝子と $n = 0, 1, 10$ における tsp による関連遺伝子の上位 8 件のリストを表 2 に示す．肥満病に対しては，推移回数が増えることによって AGRP，MC3R，MCHR1 の順位関係が正しくなっている．しかし，推移を考慮することにより結果が悪化している例もあり，どのような場合に提案手法が有効であるか，更なる分析が必要である．

4　まとめと今後の課題

本稿では，創薬支援の観点から，疾患に対して副作用の可能性の小さい特有的な遺伝子群を抽出する手法を提案した．また，CTD から作成したデータセットをもとに，順位付けアルゴリズムの精度評価を行った．

今後は，自然言語処理や機械学習を取り入れた関連用語の抽出手法の洗練化，他の語彙の使用について検討する予定である．

謝辞　本研究は科学技術振興機構バイオインフォマティクス推進事業「バイオ基幹情報資源の高準化と共有化」の支援を受けて行った．

参考文献

[1] L. Adamic, D. Wilkinson, B. Huberman, and E. Adar, “A literature based method for identifying gene-disease connections,” Proc. of the IEEE Computer Society Conference on Bioinformatics, pp. 109–117, 2002.

[2] D. Cheng, C. Knox, N. Young, P. Stothard, S. Damaraju, and D. Wishart, “PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites,” Nucleic Acids Research, 36 (Web Server issue), W399–W405, 2008.

[3] A. Özgür, T. Vu, G. Erkan, and D. Radev, “Identifying gene-disease associations using centrality on a literature mined gene-interaction network,” Bioinformatics, 24(13), i277–i285, 2008.

[4] Y. Tsuruoka, J. Tsujii, and S. Ananiadou, “FACTA: a text search engine for finding associated biomedical concepts,” Bioinformatics, 24(21), 2559–2560, 2008.

[5] S. Yu, S. Vooren, L. Tranchevent, B. De Moor, and Y. Moreau, “Comparison of vocabularies, reprensetations and ranking algorithms for gene prioritization by text mining,” Bioinformatics, 24(16), i119–i125, 2008.

統合システムを有する分散型熱物性データベースの展開

Development of network database system for thermophysical properties data equipped with integrated system

山下 雄一郎*,馬場哲也
Yuichiro YAMASHITA, Tetsuya BABA

*産業技術総合研究所
National Institute of Advanced Industrial Science and Technology
〒305-8563 茨城県つくば市梅園 1-1-1 つくば中央第 3
E-mail: yuichiro-yamashita@aist.go.jp

産業技術総合研究所ではコンテンツとシステムの所有権を分離して管理する分散型熱物性データベースが開発され、一般に公開中である。日本熱物性学会においても、同システムを利用した熱物性データベースが公開中である。これらのデータベースは、専用の閲覧ソフトウェア InetDBGV と Web ブラウザからの閲覧システム(TPDS-web/full)から利用可能である。分散型熱物性データベースプログラム 3.0.6 版からは複数のデータベースを横断的に閲覧・検索する機能が実装され、データベースを統合して操作するシステム“統合システム”として、InetDBGV を利用可能となった。本報では分散型熱物性データベースの統合システムの全容について報告する。

National institute of advanced science and technology (AIST) has developed network database system for thermophysical properties data (TPDB) and opened it on internet. In addition, Japan society for thermophysical properties launched thermophysical properties database whose system is same as AIST one. These database systems are available through installation database software “InetDBGV” and TPDS-web/full program running on web browser. From TPDB version 3.0.6, multiple concurrent connection and cross searching function were equipped with InetDBGV. Because of these functions, InetDBGV comes to play a role of integrated system.

キーワード: 分散型熱物性データベース, 標準データ, 統合システム, network database system for thermophysical properties data, standard reference data, integrated system

1.緒言

産業技術総合研究所(以降、産総研) 計測標準研究部門(NMIJ)では、前身の計量研究所の時代から、知的基盤の構築を目的として高精度・高信頼性のデータを体系的に収録した権利分散型の熱物性データベースを開発してきた[1]- [4]。権利分散型の意味はシステムの所有権とデータの所有権を分離して管理することにある。現在、産総研分散型熱物性データベースはポータルサイト[5]より利用可能である。

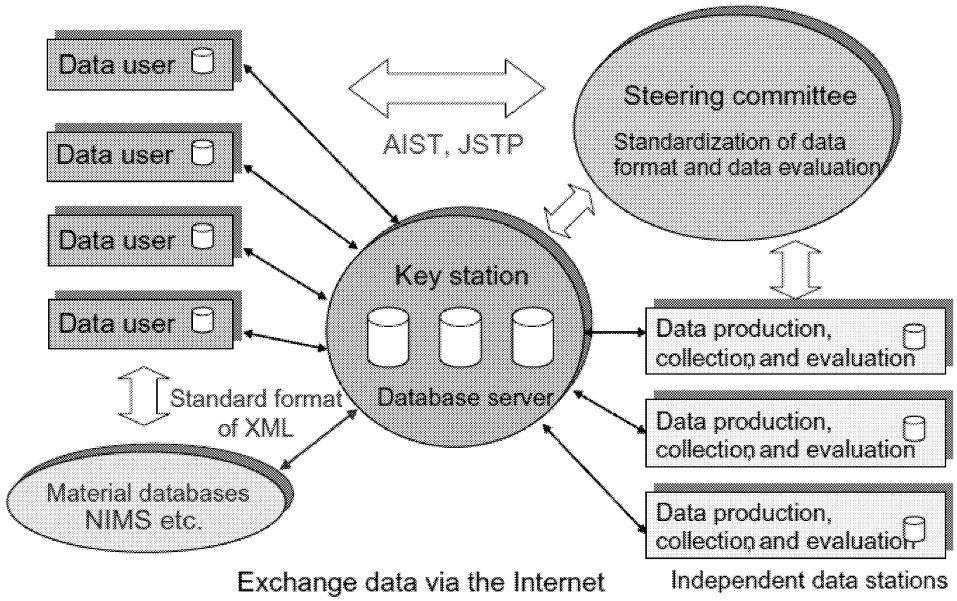


図 1　分散型熱物性データベーの連携概念

また分散型熱物性データベースのデータベースプログラムは産総研のみならず、外部機関においても利用可能である。同プログラムを利用して、日本熱物性学会(JSTP)が 2009 年 3 月より熱物性データベースを公開・運営している[6]。

分散型熱物性データベースは各研究機関と連携し、強固なデータインフラを構築することを念頭に展開されてきた。図 1 に分散型熱物性データベースの連携図を示す。Key station、運営委員会、データ生産者､データ利用者から構成される。ここで Key Station は概念であり、実際には同じプログラムで動作するデータベースサーバー群を想定している。Key Station 中のサーバー群はそれぞれが得意とする分野の熱物性データの収集・管理を行う。例えば産総研では、国家標準である熱物性計測装置で生産される標準データや､評価済みの文献データを蓄積し、データ評価のためのデータを提供することを目的としている。一方 JSTP 熱物性データベースは、学会で承認されたコンテンツ（例えば、新編熱物性ハンドブック）を収録し、学会員専用サービスとして広く熱物性データを提供する。分散型熱物性データベースの運営委員会には、JSTP 熱物性値サービス委員会が位置付けられる。そこではデータの収集・生産計画、熱物性標準フォーマット、データ評価手法の確立といった項目に関して取り決めがなされる。データ生産者は運営委員会の決定に従って、実測データを蓄積し、データ評価ガイドラインに従ってデータを評価し、Key station 中のサーバーにデータを登録する。こうした明確な役割分担の下に、サーバー毎に特色のあるコンテンツを拡充する計画である。

分散型熱物性データベース閲覧はインストール版閲覧ソフトウェア InetDBGV

Property Statistics

Property	Number of data	List of Materials
Surface tension	3719	View
Vapor pressure	2119	View
Thermal conductivity	636	View
Specific heat capacity at constant pressure	516	View
Density	488	View
Thermal diffusivity	298	View
Sound velocity	264	View
Spectral emissivity	232	View
Specific heat capacity at constant pressure per mol	193	View
Specific volume	156	View
Viscosity	144	View
Enthalpy per kg	143	View
Entropy per kg	132	View
Enthalpy per mol	116	View
・		
・		
・		
・		
Transmissivity	3	View
Yield stress	3	View
Saturated liquid density	2	View
Bubble pressure x	1	View
Extinction coefficient	1	View
Totals:	9815	

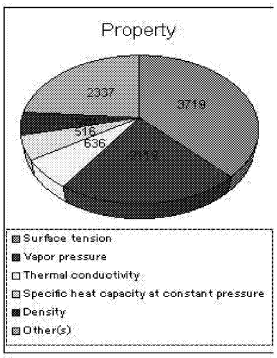


図 2　物性データ統計

ないし、Web 版閲覧ソフトウェア TPDS-web/full[7]より行う。InetDBGV は分散型熱物性データベースのポータルサイトからユーザー登録を行うことで利用可能となる。InetDBGV では大量のデータを一度に処理する機能が搭載されており、ある物質グループに対する物性データの変化を視覚的に捉えるといった使用ができる。一方、TPDS-web/full の利用にはユーザー登録は不要であり、さらに直観的で使いやすいユーザーインターフェースとなっている。こちらは特定の物質・物性をピンポイントに求めるユーザーに適している。

産総研分散型熱物性データベースプログラムバージョン 3.0.6 では、新たに 1. 複数接続機能、2. 文献検索機能、3.画像収録・表示機能が追加された。特に複数接続機能によって、複数データベースに同時にアクセスし、複数のサーバーに存在するデータを一つのクライアントソフトで処理することが可能となった。これより、データベースを統合して操作するシステム(以降、統合システム)として InetDBGV を位置付け、本報では統合システムとしての InetDBGV とその新機能、活用法について報告する。

2.分散型熱物性データベース

分散型熱物性データベースはインターネットを利用したサーバー・クライアント型のデータベースである。2009 年 4 月 10 日現在、産総研分散型熱物性データベースではデータベース登録ユーザーに 9815 件の物性データを公開している。物

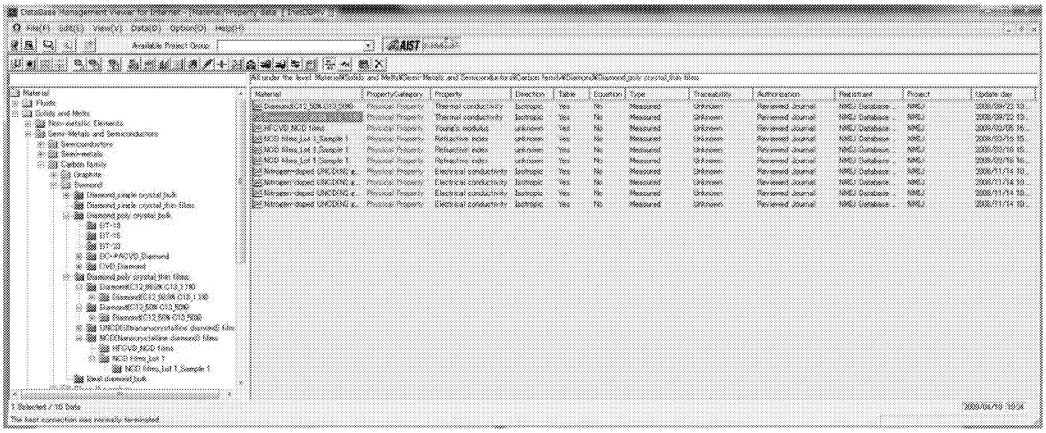


図 3　統合システム InetDBGV スクリーンショット

性データ毎の内訳を図 2 に示す。JST が所管していたKelvinデータベースが産総研に移管された経緯もあり、表面張力、蒸気圧のデータが多く、次いで熱伝導率、比熱容量となっている。産総研ではデータの収録に継続的に取り組んでおり、今後さらに充実する見込みである。

分散型熱物性データベースでは、物質データ、物性データは分離して管理し、リンクテーブルを用いてそれらを結びつけている。物質データは細かく分類され、データベースに体系的に整備されて収録される。分類の基準として、大分類では学協会や産業分野、中分類では材料の特性や構造、小分類では計測時の試料情報となっている。大分類から小分類にしたがって、物質が同定されていく仕組みである。詳しくは参考文献[3]を参照されたい。一方、物性データには、物性名(Property)、物性の方向(Direction)、表データ(Table)と式データ(Equation)の有無、データ取得情報(Type)、トレーサビリティの状態(Traceability)、認証状態(Authorization)、データ登録者(Registrant)、データ取得プロジェクト(Project)、データ更新日(Update date)が基本情報として備わる。これらの他に、文献情報、不確かさ情報、実測データ解析手法、装置情報、計測規格が追加情報として記述可能である。

3.統合システム InetDBGV と新機能

インストール版閲覧ソフトウェア InetDBGV の基本画面を図 3 に示す。基本画面は画面左の物質ツリーと画面右の物性データリストから構成される。InetDBGV では階層構造を利用することで分類された物質データを表現する。一方、物性データリストには物性データ基本情報のみが表示され、物性データ追加情報は右クリックメニューから閲覧することができる。右クリックメニューからは物性データのグラフ化なども行える。

3.1 複数接続機能

緒言で述べたように、分散型熱物性データベースは同じプログラムで動作するサーバー群で Key station を構成する。この時、サーバー間のインターネットを介した通信は発生しない。分散型熱物性デ

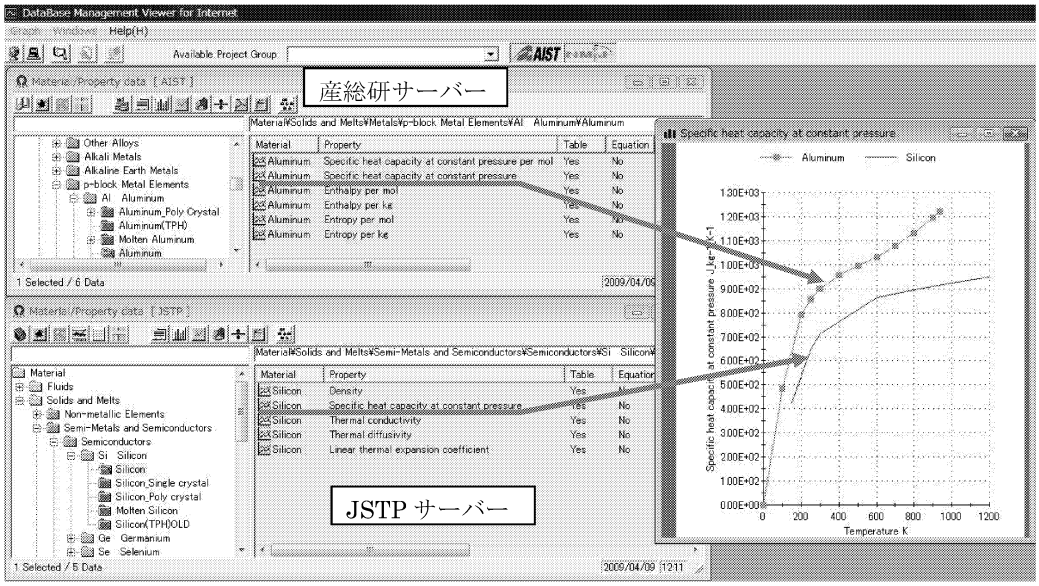


図 4　複数接続機能と同一グラフ上への描画

ータベースでは、これらのサーバーに収録されたコンテンツを有機的に結び付けるために、クライアントソフトウェアである InetDBGV から複数サーバーに対して同時かつ横断的にアクセスすることで対応している。各サーバーへのアクセスとデータ閲覧許可のコントロールは、InetDBGV が備えるユーザー名とパスワードによるログイン機能を利用する。この為、ユーザーは初期設定で Key station 中のサーバー群に自由にログインできるのではなく、事前にサーバー管理者にログインアカウントの申請を行い、アカウントとサーバー接続先 URL を取得する必要がある。複数接続機能を開発したことで、図 4 のように産総研と JSTP の 2 つのサーバーからデータを取得し、1 つのグラフ上に描画することが可能となった。

3.2 検索機能

3.2.1 文献検索

分散型熱物性データベースは物性データを収録したファクトデータベースである一方で、物性データには文献情報を付帯させており、文献データベースとしての側面も持つ。バージョン 3.0.6 からは、文献タイトル(論文名、雑誌名、書籍名)、巻、出版年、開始ページ、終了ページ、著者での検索が可能となった。ここで、文献タイトルに関しては、部分一致検索である。また、文献検索機能は、物性検索(物性名、物性値範囲、変数値範囲)、物質名検索、付帯情報検索(CAS No.、化学式、材料形状、RM/CRM コード、結晶構造、材料応用分野)などと組み合わせて AND 検索を行うことができる。

3.2.2 横断検索

複数サーバーに接続した場合、任意のサーバーに対して横断的に検索をかけることができる。図 5 に横断検索を行った

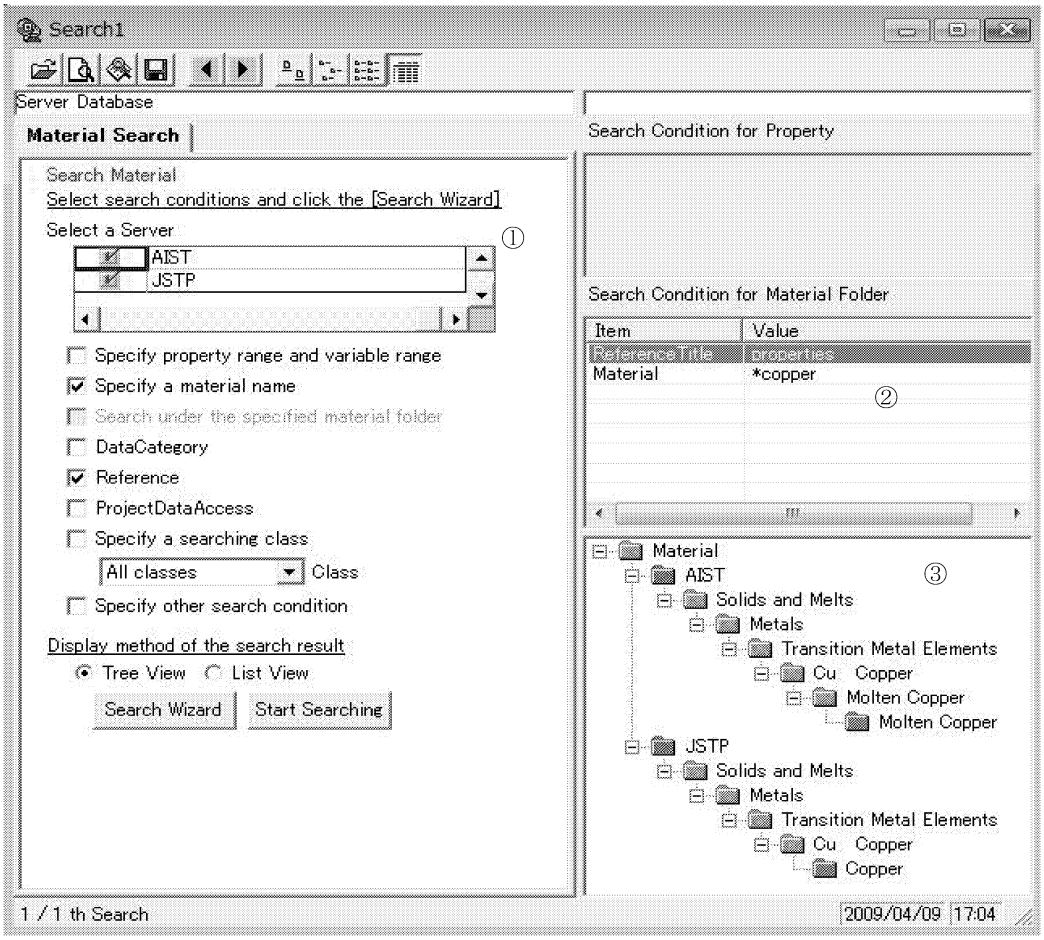


図 5 横断検索結果

結果のスクリーンショットを示す。図 4 中①の部分で検索対象サーバーを選択することができる。現状で 5 つのサーバーまで同時に検索可能である。検索項目は②に表示され、図の例では、文献タイトルに「Properties」が含まれ、かつ物質名が Copper で終わるものを対象としている。検索結果は図 4 中の③部分に表示され、検索対象サーバーごとに検索結果が分かるようになっている。検索結果の表示部分はツリー表示、リスト表示の両方から選択可能である。

4.統合システムの活用

3.1 節で述べた複数接続機能の具体的な活用方法としては、標準データを利用したデータ評価が挙げられる。従来であれば、標準データをテキスト論文中の表をタイプし直したり、グラフから値を読み取ったりしていたが、そうした作業は手間がかかる上に、人的ミスが混入する可能性を含む。分散型熱物性データベースでは、標準データを集中的に収録したデータ評価用サーバーを準備すれば、ユーザーは同サーバーからデジタル情報で

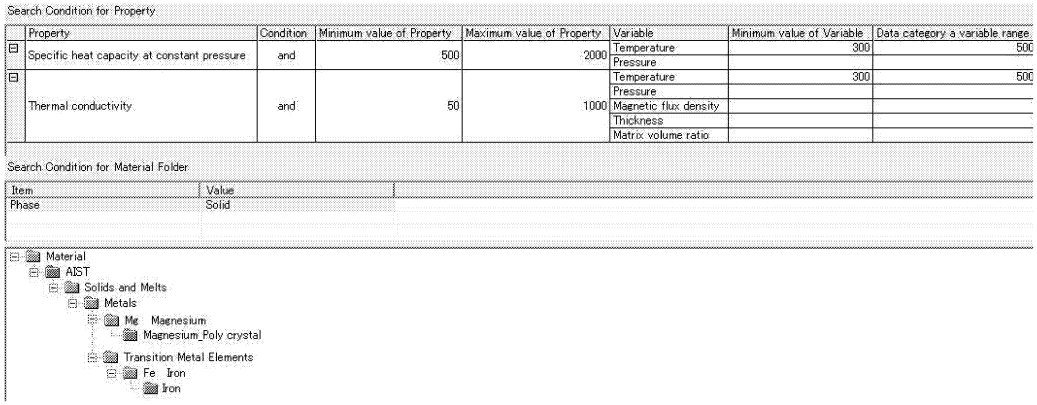


図 6　物性値範囲指定検索

標準データを受け取ることができる。ユーザーはデータ評価用サーバーと収録先サーバーの両方にアクセスすることで、InetDBGV 上でデータ比較を行ことができる。その際、グラフの重ね描き機能を利用すれば、視覚的にデータ評価を行える。もう一つの活用例として、新規材料の特性評価が挙げられる。新規材料の有効性は特性データを過去のデータと比較して、初めて評価される。InetDBGV では複数のサーバーから網羅的にデータを取得し、大量のデータを一度にグラフ化する機能に優れており、新規材料がどういった特徴を有し、どういった材料と競合するかなど、客観的に材料特性が理解できる。こうした大量データに裏付けられた情報は、新規材料の開発方針へとフィードバックでき、材料製作から特性評価を経て、次の材料試作方針を決定するサイクルの迅速化に貢献できる。

3.2 節の検索機能については、ファクトデータと文献データの両方を備えることで、それぞれの特徴を融合した検索サービスが提供可能である。例えば、特定の物性・温度領域・材料について計測実績があるかどうか、計測実績があれば、その文献情報まで知りたい場合がそうである。今日の Google 等の検索エンジンはキーワード検索が主体であり、物性範囲・材料を指定した検索には対応していない。そうした場合、検索結果に表示されたリンク先を丁寧に調査するほかない。一方 InetDBGV では、物性検索と物質検索を同時に利用することで物性データの絞り込みが可能であり、その物性データから引用文献を逆引きすることができる。例として、図 6 に物性範囲指定検索の結果を示す。図 6 は、熱伝導率が 50～2000W/mK の範囲かつ定圧比熱容量が 1000~2000J/kgK の範囲にあり、温度が 300～500K の範囲で範囲指定を行い、さらに固体状態に限って検索を行った結果である。検索結果として、多結晶マグネシウムと鉄がヒットしている。検索結果のツリーに表示される物質名部分をダブルクリックすることで、物性データウィンドウに遷移し、該当物性データの文献情報を閲覧することができる。さらに分散型熱物性データベースでは、図 7 のように物性データにオンラインジャーナルの URL も付帯しており、ユーザーが該当のオンライジャーナルへのアクセス権を

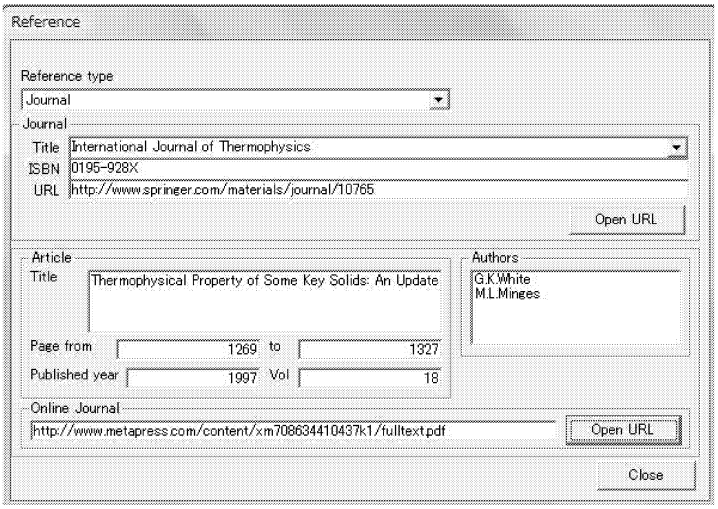


図 7 物性データに付帯する文献情報

持っていれば、ワンクリックで該当文献の PDF までたどり着くことが可能である。

5. 結言

本報告では、分散型熱物性データベースにおいて統合システムとして InetDBGV を位置づけ、その新機能と活用法について述べた。今後、各機関と提携しコンテンツに独自色を持ったサーバーを増やすことで、情報利用の利便性が高まり、最終的にユーザーによりよいサービスを提供できると考えられる。また、この統合システムは完成されたものでなく、今後ユーザーのニーズを受けて数々のアップデートを行っていく計画である。

参考文献

[1] A.Ono, T.Baba, K.Fujii: “Traceable measurements and data of thermophysical properties for solid materials: a review”, Meas. Sci. Tech. 12, pp.2023-2030, 2001.

[2] 馬場哲也:「科学技術を支える知的基盤:熱物性データベース」, 第 25 回日本熱物性シンポジウム予稿集, 長野市, pp.89-91, 2004 年.

[3] 馬場哲也:「分散型熱物性データベースの開発」, 熱物性, 18, pp.136-142, 2004.

[4] 馬場哲也:「物性データベースの本格研究 熱物性データの生産から供給へのハブ構築」, AIST Today, 7, pp.18-19, 2007.

[5] 分散型熱物性データベース: http://riodb.ibase.aist.go.jp/TPDB/DBGVsupport/.

[6] 日本熱物性学会熱物性データベース: http://jstp.mech.nagasaki-u.ac.jp/TPDS-web/.

[7] TPDS-web/full: http://riodb.ibase.aist.go.jp/TPDB/AJAX/.

マテリアルリスク指標データベースの設計と活用

Design and Application of Database for Materials-Risk-Index

芳須　弘[1*], 原田幸明[1], 藤田充苗[1]
Hiroshi YOSHIZU, Kohmei HALADA and Mitsutane FUJITA

*1　物質・材料研究機構
National Institute for Materials Science
〒305-0047　茨城県つくば市千現 1-2-1
E-mail: YOSHIZU.Hiroshi@nims.go.jp

地球環境問題を考慮した材料の持続的供給を行うために，材料設計や材料選択において必要な情報を検討し，マテリアルリスク指標が提案されている．その指標は材料の持続的供給，環境変動および毒性のリスク指標からなっている．本報告では，これら3種類のリスク指標に関する情報を蓄積できるデータベースの概念設計について検討した．さらに，XML で記述されたそのデータベースと材料特性データベースとを連携した場合，耐熱合金の供給持続性や環境を配慮した材料選択の方針を紹介した．

In order to do the sustainable supply of materials considered environmental problems in Earth, the necessary information in materials design and selection is discussed and the Materials-Risk-Index is proposed.　The index consists of three kinds of risk index of the sustainable supply, environmental variation and toxicity of materials.　In this report, the conceptual design of the database to store the information about three risk indexes was considered. If the database and mechanical properties databases described by XML were cooperated, it was introduced that the sustainability of supply and a guide of new material design in heat resistance alloys.

キーワード：持続的社会，マテリアルリスク指標，データベース，XML，材料選択，
Sustainable society, Materials-Risk-Index, Database, XML, Materials selection

1　はじめに

地球環境問題を考慮した持続可能な社会を形成することが，地球温暖化などの環境破壊解決に繋がると考えられてきた．持続的社会の実現を念頭に，構造物や機器に用いる材料を選択する場合，材料特性の他にも構成する物質の持続供給可能性，製造過程での汚染物質の排出量，使用時の毒性などの環境適合性に対する考慮が社会的に一段と要求されている．これまで，エコマテリアルの観点から材料設計・選択に関する提案がなされている．この状況下，材料設計・選択には物質・

材料の持続供給，製造時の汚染物質排出あるいは安全性を評価するマテリアルリスク指標（Materials Risc Index）データベース（以下，データベースを DB と略）が不可欠であり，その構築へ向けた設計等を検討している[1].

一方，インターネット上の Web サイトには多くの情報が提供されており，耐熱材料に関して必要な情報はサーチエンジンで検索することによって利用することができる．しかし，いくつかの情報を組み合わせて新たな知見を得るには，利用者の目的に応じて必要な情報を同時にあるいは順序良く取得する必要がある．それには，マテリアルリスク指標に関する情報の XML 化が不可欠である．この XML 化は，既に提案されている材料特性を記述する MatML[2]，MatDB[3]などとも連携し，情報の共有を容易に図れるものでなければならない.

そこで，物質・材料のマテリアルリスク指標 DB を設計するにあたり，データ構造，必要な機能，XML 化および構築された際の活用について検討し，問題点を把握することを目的とした．マテリアルリスク指標 DB の活用については，高温機器の構成材料に用いられる耐熱材料の供給持続性および環境を配慮した材料選択の方針を検討した.

2　マテリアルリスク指標

地球環境問題を考慮した持続可能な社会を形成するには，材料分野の立場から，材料の資源採掘，製造，使用，そして廃棄までの過程でどのような情報の蓄積が必要かを検討する．すなわち，対象の材料を使用することが持続供給可能であるか否かを判定するのに必要な情報である．持続可能な社会を念頭において，材料の選択や設計あるいは材料製造工程の変更などを製品設計者が利用できる簡易な指標を社会が共有しなければならない．この共有できる指標は，社会への影響度を加味したもので，マテリアルリスク指標に相当する．その指標は，物質あるいは材料の採掘，精錬，加工の製造過程を経て，材料が使用された後リサイクルや廃棄までにかかわる情報を簡潔に表現するものでなければならない．マテリアルリスク指標は，図１に示すような 3 種類のリスクがあり，機器設計などに使用される材料が持続的社会を形成するか否かの判定に少なくとも必要となる．すなわち，材料を構成する物質の持続的供給の指標を示す”持続的供給リスク”，物質を地殻から採掘し，各種の機器に使用する材料となるまでに地球環境を変える指標を示す”環境変動リスク”および製造された材料が使用中の毒性の指標を示す”毒性リスク”の 3 種類の情報である.

持続的供給リスク指標は，物質の必要量や供給量と社会的影響度で表わされる．社会的影響度としては，物質コストや関与物質総量，資源枯渇の程度などが目安となる.

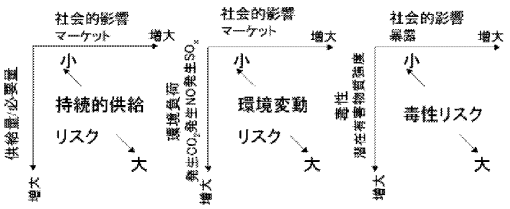


図 1　持続的供給，環境変動，毒性の 3 種類のリスク指標からなるマテリアルリスク指標

環境変動リスク指標は，物質あるいは材料を製造する際の環境負荷排出量（CO2，NOx，SOx），消費エネルギー量，さらに製造に関与する物質総量（TMR））と社会的影響度で表わされる．社会的影響度としては，温暖化指数，材料コスト，などの数値が目安となる．

毒性リスク指標は，人工的に製造された物質あるいは材料を使用する際に，それらが持つ潜在有害物質強度と社会的影響度で表わされる．社会的影響度は対象の物質や材料の使用量すなわち暴露量や形状に依存して決められる．

3 種類のマテリアルリスク指標の情報が，材料に係わる研究者や技術者に共有され，機器設計における材料選択や材料設計に反映される必要がある．しかし，3 種類のリスク指標を決める社会的影響の度合いを決めうる数値項目が社会で共通に認識さているものはない．そこで，社会的影響度を表現する項目を種々選択でき，最適の数値項目を共通に認識できるシステムとして，まずマテリアルリスク指標 DB（MRIDB)の構築を検討する．

3 マテリアルリスク指標データベース（MRIDB)

3.1 概要

マテリアルリスク指標に関する情報を蓄積した DB は，図 2 に示すように持続供給，環境変動，毒性リスク指標からなる 3 種類のサブ DB から構成される．これらの DB が十分活用され，持続可能な社会の実現への指針を提供できる材料問題解決用セマンティック Web とするには，材料特性 DB や材料用途 DB などとの連携をはじめとし，多くの Web 上に提供される情報との連携共有が不可欠である．それには，サブ DB のデータが容易に共有できるように指標関連のデータ項目を XML 化し，データの共有を容易にする必要がある．

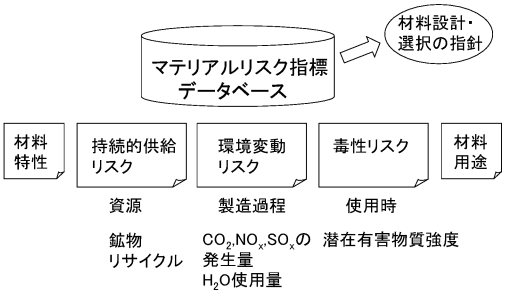


図 2　マテリアルリスク指標データベースの概念図

3.2 各リスク指標のデータ項目

以下に持続供給，環境変動および毒性リスク指標の 3 種類のサブ DB に蓄積する大まかなデータ項目について述べる．

3.2.1 持続的供給リスク指標

持続的供給リスク指標 DB に記述するデータ項目は図 3 に示したような項目が挙げられる．この DB では，物質（資源）をメインとして，埋蔵量，生産量あるいは年間使用量などの物質量に関する項目やその物質の生産に関与する TMR（関与物質総量，Total Materials Requirement）[4]がある．また，資源消費に関するパラメータである資源枯渇加速係数がデータ項目としてあり，これは資源の市場価格（平均価格）と資源の消費ベースの影響を取り入れたものである[5-6]．これらのデータ項目のデータにより物質の持続的供給

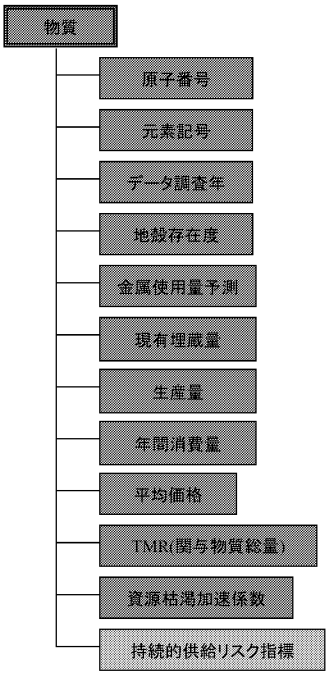


図 3　持続供給リスクに関するデータ項目

リスク指標を算定する.

3.2.2 環境変動リスク指標

環境変動リスク指標 DB は地球上の様々な物質を使用して製品製造までの一連の過程における地球環境に及ぼす情報を記述する．図 4 はこの DB に記述するデータ項目を示す．ここでは材料分野の観点から材料をメインとして材料を構成する成分（合金元素など）を記述し，材料の製造法や製造時に発生する CO_2 などの発生量を環境負荷データとして記述し，さらに製造時の水使用量や水質汚染などの環境に関わる項目を記述する．また, 材料を構成する成分の採掘から製造, 使用そして廃棄までのエネルギーの流れを物量フロー, TMR フローとして記述する．これらの情報により環境変動リスクの指標を算定する.

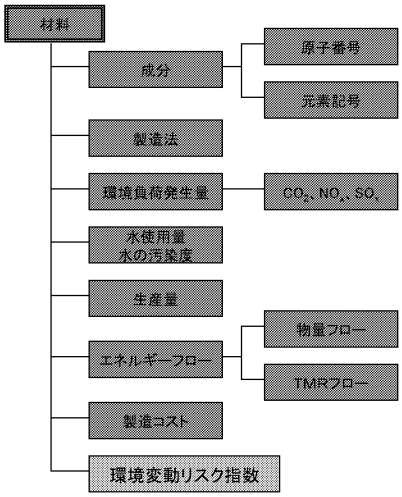


図 4　環境変動リスクに関するデータ項目

3.2.3 毒性リスク指標

毒性リスク指標 DB は対象となる物質をメインとして，対象物質が単一の化学物質の場合と製品に含有されている場合に分類して記述する．図 5 は毒性リスク指標 DB に記述するデータ項目を示す．図に示したデータ項目は，MSDS（物質

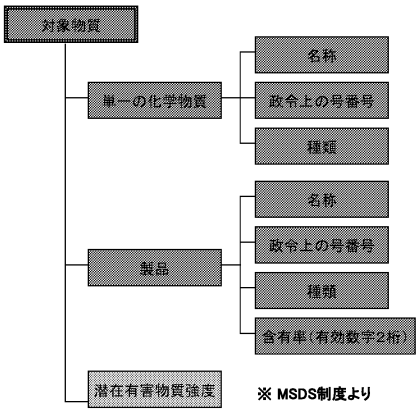


図 5　毒性リスクに関するデータ項目

安全データシート，Material Safety Data Sheet）と呼ばれ，国内規格として JIS Z 7250, 国際規格として ISO11014-1 にその記載内容が標準化されており，この MSDS の記載内容から参照した[7-8]．MSDS をもとに，この DB に記述するデータ項目は単一，製品に共通する項目として物質の名称，政令上の号番号，そして物質の種類を記述する．製品の場合はその物質の含有率を記述する．これらの情報をもとに潜在有害物質強度とともに毒性リスク指標を算定する．

3.3 必要機能

サブ DB のそれぞれの蓄積データは，政府の種々の専門機関から Web 上に提供されている各種の統計情報を整理・加工し，容易に活用できる形式で収録できる機能が必要となる．

Web 上で提供されるマテリアルリスクに関する統計情報を整理，蓄積し，その情報の共有，データの変換および関連数式との連携を容易にするために，3 種類のサブ DB に必要なデータと機能を検討する．3 種類のサブ DB に蓄積した情報から，社会的影響度を表すと考えられる数値項目を任意選択でき，それぞれのリスクの最適の指標を算出し，指標の数値化可能な機能が必要である．

情報の連携共有には，MRIDB 関係の情報の XML 化が不可欠である．マテリアルリスク指標の XML 化は，材料特性の情報を記述する MatML や MatDB との連携，さらに数式の記述用の MathML[9] など異種分野の ML（マークアップ言語）と連携して情報の共有が容易であることが要求される．

図 6 は持続的供給リスク指標 DB に関するデータ項目を XML 形式により記述した例を示す．この DB では，前述したように物質をメインとして物質の埋蔵量，使用量あるいはその物質に関与する情報などを記述する．

```
<?xml version="1.0" encoding="Shift_JIS" ?>  <!DOCTYPE
material (View Source for full doctype...)> 
<持続的供給リスク>
    <物質>
        <原子番号>3</原子番号>
        <元素記号>Li</元素記号>
        <データ調査年>xxxx/xx/xx</データ調査年>
        <地殻存在度>13</地殻存在度>
        <年間消費量 unit="ton/year">234300</年間消費量>
        <現有埋蔵量 unit="ton">11000000</現有埋蔵量>
        <平均価格 unit="$/kg">3.89</平均価格>
        <TMR>1500</TMR>
        <資源枯渇加速度>1800</資源枯渇加速度>
        <資源指標>600</資源指標>
        <持続的供給リスク指標>?</持続的供給リスク指標>
    </物質>
                    ⟩
    <物質>
        <原子番号>92</原子番号>
        <元素記号>U</元素記号>
        <データ調査年>xxxx/xx/xx</データ調査年>
        <地殻存在度>0.91</地殻存在度>
        <年間消費量 unit="ton/year">66529</年間消費量>
        <現有埋蔵量 unit="ton">458000</現有埋蔵量>
        <平均価格 unit="$/kg">10</平均価格>
        <TMR>22000</TMR>
        <資源枯渇加速度>32000</資源枯渇加速度>
        <資源指標>10000</資源指標>
        <持続的供給リスク指標>?</持続的供給リスク指標>
    </物質>
</持続的供給リスク>
```

図 6　持続的供給リスク指標 DB に関する XML 記述

4 マテリアルリスク指標データベースの活用

MRIDB が整備されたことを前提として，各リスク指標の活用について持続的供給リスクと環境変動リスクの 2 つのリスク指標の活用を検討した結果の例を以下に示す．

4.1 耐熱材料の持続的供給リスク指標

耐熱材料の種々の合金系の持続的供給リスク指標を材料の構成元素と価格や公

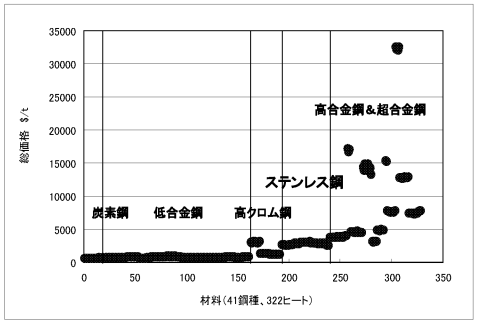


図 7　材料製造に関する構成合金元素の総価格

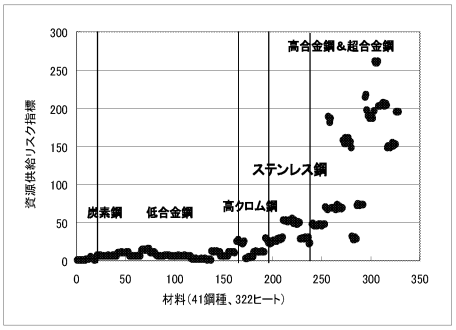


図 8　材料製造に関する資源供給リスク指標

表されている資源供給リスク指標を示す数値データなどを社会的影響度として算出した．その結果を図 7 および図 8 に示す．

図 7 は（独）物質・材料研究機構のクリープデータシート 41 鋼種（322 ヒート）[10]の各 1 トン当りについて，持続的供給リスク指標をもとに各材料を構成する主要合金元素を対象としてそれぞれの平均価格（市場価格）を参照し，その総価格を算出し示したものである．この図から高額，すなわち採掘，使用，リサイクルと一連の流れの中で循環する量が少ない合金元素を多く含有する高合金鋼や超合金は総価格が高額を示めす．また，炭素鋼のように生産量やストック量のある鉄が多くを占めている材料は低価格を示す．

一方，これらの材料を構成する合金元素の資源供給におけるリスクについての結果を図 8 に示す．資源供給リスクについても価格と同様に主成分の現有量，生産量に依存し，高合金や超合金などのように添加する合金元素の数や添加量も影響する．このように，超合金系の耐熱材料の持続的供給リスク指標が高いことがわかる．

以上のように，材料分野の観点からリスク指標 DB 活用として材料と価格や資源枯渇リスクとの関係を調べた結果を示した．材料を製造する上で考慮すべきこととして材料を構成する合金元素の流通，価格や将来の資源枯渇を念頭に入れておく必要がある．

ここまで各リスク指標の活用について示したが，以下は異種分野の情報との連携について検討した結果を述べる．

4.2 クリープ特性 DB と環境変動リスク指標サブデータベースとの連携

ここでは 3 つのリスク指標の中からから環境変動リスクを取り上げ，材料特性分野の情報との連携を行い，両者の関係を調べた．ここで用いた材料は同じくクリープデータシートの 41 鋼種である 環境変動リスクと材料特性の関係をみるために，環境変動リスクについては各材料の主要成分をもとに材料製造時（溶解，圧延など）に排出される環境負荷データ（CO_2，SOx，および NOx）を算出した．

一方の材料特性については各材料（各材料に複数ヒート）の長時間強度データ

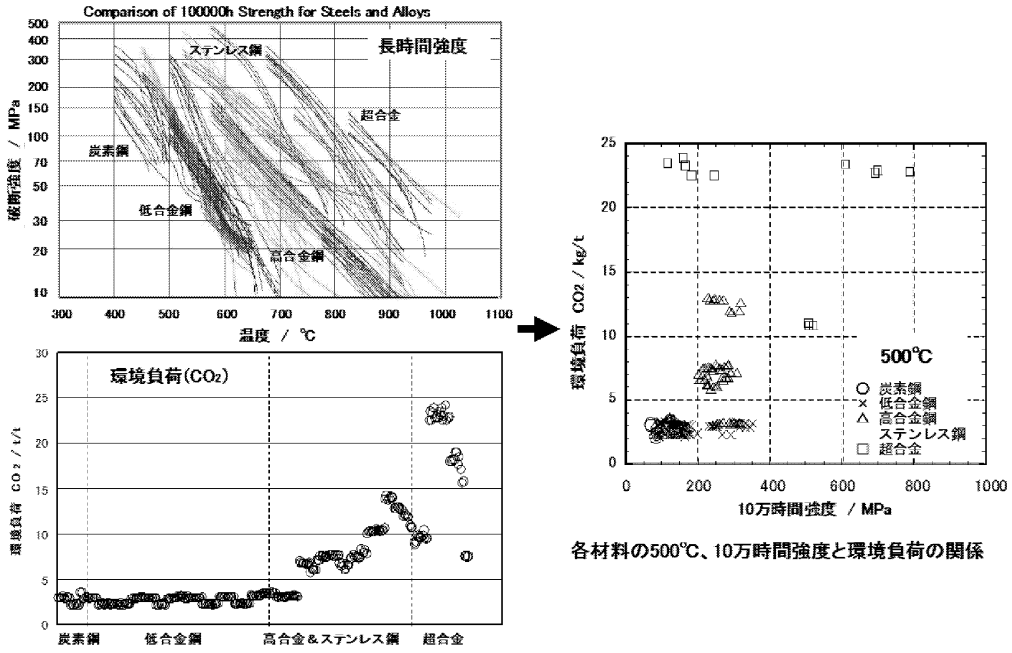


図 9　各材料の長時間強度と環境負荷の関係

（10 万時間）を算出した．これらのデータを用いて両者の関係を調べた．その結果を図 9 に示す．図 9 の左上に各材料の長時間強度（10 万時間強度）を示し，左下には各材料の環境負荷（CO_2 の結果を抜粋）を示した．長時間強度は各材料の用途や適用される温度域が異なるため炭素鋼の 450℃前後から超合金の 950℃前後まで幅広くなっている．環境負荷はやはり合金元素の数や製造法などが影響し，合金元素の数が少ない炭素鋼や添加量が少ない低合金鋼などは環境負荷が低く環境への影響が少ない．一方，鉄ベースと異なりニッケルやコバルトをベースとする超合金は合金元素の数や合金元素の製造過程（採掘から製品まで）が影響し環境負荷が大きくなっている．

これら両者の情報をまとめてその関係を調べた結果を図 9 の右に示す．この図は各材料の 500℃における長時間強度と環境負荷（CO_2)の両者の関係をみたものである．環境負荷（CO_2)は各材料 1 トンを得るために排出される量を示す．また，超合金についてはその適用温度が高いため低温の長時間強度の算出が不可能であり一部を除外した．図の結果から，各材料に必要な長時間強度を得るためには地球環境にこれだけの影響を与えていることを示しており，材料製造の際にはこのような結果を考慮する必要がある．

5　おわりに

本報告では，材料に関わるマテリアルリスク指標をもとに，その DB 化と DB の活用を検討した．その結果は以下の通りである．

1) マテリアルリスク指標 DB は，持続的供給，環境変動および毒性の 3 種類のサブ DB から構成され，それぞれのサブ DB で蓄積すべき主要データ項目を明らかにした．
2) MRIDB は，蓄積数値データの収集や利用のためには収録情報を XML で記述することが不可欠である．
3) MRIDB の活用例として，持続供給サブ DB と材料特性 DB との連携で，各種の耐熱材料の持続性を明らかにした．
4) 環境変動サブ DB と材料特性 DB との連携では，環境負荷ガス(CO_2)と各種耐熱材料のクリープ特性も関係から，環境負荷を低減する材料選択の方向を示した．

本研究は，日本学術振興会科学研究費補助金の支援をもとに検討を進めている．

参考文献

[1] 日本学術振興会，科学研究費補助金（基盤研究 B)，「環境材料設計基準に資するマテリアルリスク指標」，2008
[2] MatML, http://www.matml.org/（2009 年 2 月参照）
[3] MatDB, インターネット版先進金属材料の各種材料特性ファクトデータベース, http://www.nims.go.jp/vamas_twa10/AMM_DB/index.html（2009 年 2 月参照）
[4] 中島謙一他：「関与物質総量（TMR）の算定，Journal of Life Cycle Assessment, Japan, Vol.2, No.2, pp.152-158, 2006.
[5] 原田幸明：「希少資源・元素の現状」, まてりあ, vol.46, No.8, pp.543-548, 2007.
[6] 製品環境情報提供システム事務局：「資源枯渇性特性化係数について」, 2005.
[7] 経済産業省化学物質管理課 HP：http://www.meti.go.jp/policy/chemical_management/law/msds/index.html（2009 年 3 月参照）
[8] 独立行政法人製品評価技術基盤機構 HP（http://www.prtr.nite.go.jp/）（2009 年 3 月参照）
[9] MathML， http://www.w3c.org/Math/，（2009 年 2 月参照）
[10] 独立行政法人物質・材料研究機構, クリープデータシート, No.1~No.41.

テキスト批評の計量化に向けて－書評の計量分析－

Towards a Quantitative Research Approach to Literary Review: Quantitative Analysis of Book Reviews

村井源[1]*，徃住彰文[1]
Hajime MURAI, Akifumi TOKOSUMI

*1 東京工業大学大学院社会理工学研究科
Graduate School of Decision Science and Technology, Tokyo Institute of Technology
〒152-8552 東京都目黒区大岡山2-12-1-W9-48
E-mail: {h_murai, akt}@valdes.titech.ac.jp

本研究では，テキスト批評における中心的概念とそれらの関係を調べるため，ジャンルによる中心的概念の相違とそれらの構造を計量的に抽出した．データとしては，「Web本の雑誌」で公開されている，新刊書籍への全書評（960件，5521冊分）を対象とし，全批評テキストと各ジャンルのみの批評テキストでの出現頻度の差異からジャンルに特徴的な語彙を抽出した．また，出現頻度の高い語彙の共起ネットワーク化を行って，テキスト批評で用いられる概念の構造を可視化した．

In order to examine central concepts and their relationships within review texts, central concepts and their structures for each text genre are extracted and compared. The analyzed data consists totally of reviews for "Web Magazine of Books" (960 review articles covering approximately 5,521 books). The characteristic vocabularies of each genre are extracted according to frequency differences with the total texts and the genre texts. Based on co-occurrence frequencies, the main vocabularies are visualized as a network.

キーワード: 書評，ネットワーク，共起単語
Book review, network, co-occurrence

1 はじめに

テキスト読解における情報処理過程は，主に心理学や認知科学の分野において，その解明が進められてきた．文字を単語として認識し，単語を文節やセンテンスレベルに構造化し，センテンス間の修辞関係や論理的関係を認識するという多段階の並行的なメカニズムは非常に複雑である．テキストの読解にはコンテキストの理解の問題も含まれるため，現在まで多くの研究が進められているものの未だ

解明されていない点も少なくない.

一方で，テキスト読解の後に続く評価と批評の認知プロセスもまた複雑である.テキスト読解後の内容の把握に関しては,教育工学でconceptual map[1]などを用いた研究が進展してきているものの，テキスト批評の認知メカニズムは，何をもって批評と言いうるかという根源的な問題も含めて端緒についたばかりである[2].

本研究においては，未だ未解明の点の多い批評の認知メカニズムを探るため,最初のステップとして批評文を収集してそれらの計量的な分析を行い，批評文の全体的な概要を語彙の側面から検討する.

また，これらの計量的特徴の批評対象となるテキストのジャンルごとの差異を抽出し，ネットワーク形状に可視化して表示を行う．また，ネットワーク分析の手法を援用して，各ジャンルの批評文における重要な概念を比較する.

2 対象となる批評文テキスト

本研究において対象とする批評文テキストのデータは，「本の雑誌」における新刊書紹介コーナーの書評である．この書評のテキストはWeb上にバックナンバーが公開されている[3].一つの書評あたり5冊程度の新刊書を紹介するため,一つの書籍あたりの批評文としての量は多くないが，重複と文字化け等を除いた960書評中に5521冊分の批評テキストが含められている．一つの書籍に対して複数のジャンルがつけられる場合があるため（歴史エンターテイメントなど）単独と重複に分けてそれぞれのジャンル数表1に示した．なお，書評テキストは26人の批評家によって交代で記されている.

表1 ジャンルごとの対象書籍数

	単独	重複	合計
エンターテイメント	1468	132	1600
歴史	189	33	222
純文学	417	13	430
恋愛	17	127	144
ミステリー	1159	9	1168
SF	746	0	746
ノンフィクション	1244	1	1245
未分類	106	0	106
合計	5346	315	5661

3 批評文の頻出語彙ネットワーク

テキストに含まれる概念構造としては,内容的側面と感性的側面の両面が考えられるが[4],本研究では頻出名詞に基づいて，批評文テキストでの内容的概念構造を計量的に分析するためネットワーク状に可視化する．ネットワーク化によって概念間の関連性が視覚的に把握できるとともに，ネットワーク分析手法を用いることで種々の計量的分析が可能となる.

本研究においては，頻出語彙の影響力を可視化するため，ネットワーク図でのノードのフォントサイズは，単語の出現頻度の二乗根に比例させている．また，孤立ノードは割愛した.

図１から図７に，ジャンルごとの頻出共起単語間の関係性をネットワーク図として表示する．以下の図では，可視性のため各ジャンルでの頻出名詞の上位20に対してのみ,名詞間の共起頻度を計量し,共起頻度上位約50までをエッジとして描画している.

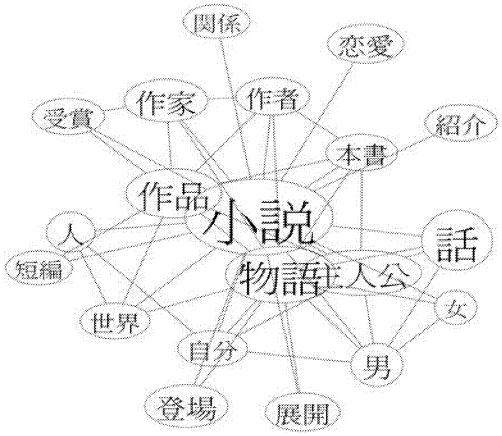


図1 エンターテイメントの語彙ネットワーク

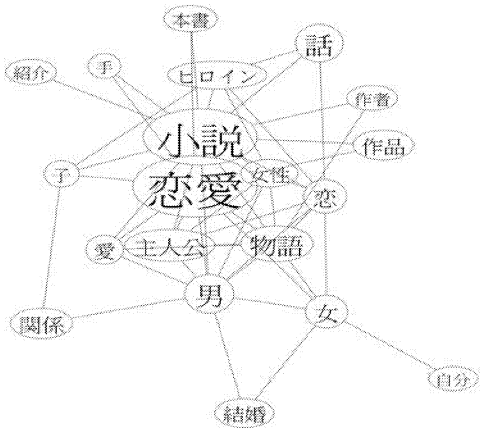


図4 恋愛の語彙ネットワーク

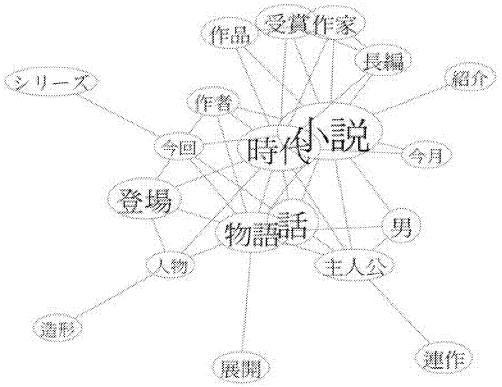


図2 歴史の語彙ネットワーク

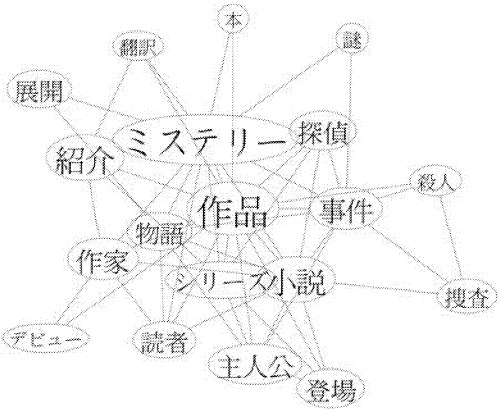


図5 ミステリーの語彙ネットワーク

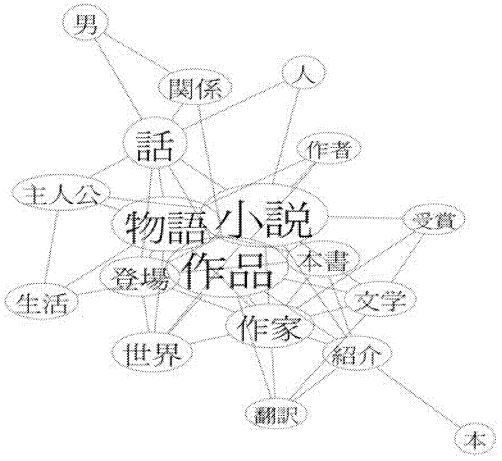


図3 純文学の語彙ネットワーク

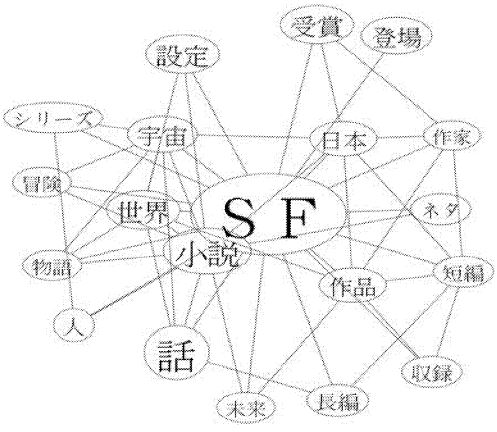


図6 SFの語彙ネットワーク

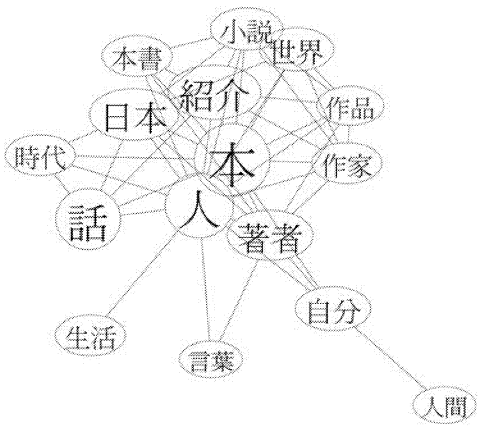


図7　ノンフィクションの語彙ネットワーク

当然のことであるが，フィクションとノンフィクションはネットワークの構成が大きく異なっている．フィクションの各ジャンルでは，「小説」「作品」「物語」など共通する語彙が含まれており，フィクション作品一般の評論に共通する書評の中心的話題が存在するとわかる．

一方で，フィクションの各ジャンルにおいて特徴的な語彙もあり，「恋愛」，「ミステリー」，「SF」などのように，ジャンルの名称が中心に来たり，他にもジャンルとの関連性が強いと考えられる，歴史ジャンルでの「時代」，ミステリージャンルでの「事件」や「探偵」，SFのジャンルでの「未来」や「宇宙」などの単語が複数含まれていたりもする．

これらの７つのジャンル中で，純文学とエンターテイメントは，頻出する語彙とそれらの関係性が非常に似通っている．純文学と娯楽小説は質的に大きく異なるものと一般的に考えられるが，これらを対象とした批評を行う場合，その大まかな切り口となる概念は似通っていることがネットワーク図より示唆される．

4　ジャンルに特徴的なキーワード

頻出語彙とそれらの共起頻度に基づく分析では，「小説」「作品」「物語」「主人公」など，全体にほぼ共通する語彙が多く，逆にジャンル固有の語彙はさほど目立たない．そこで，ジャンルに特徴的な語彙を抜き出すため，キーワード抽出において一般的に用いられるIDF分析を変形した，CIDF (Comparative IDF)を指標として用いる．

CIDFは,全テキストにおける単語のidf値idf_{all}と,特定ジャンルのテキストのみでのidf値idf_{genre}を計算し，そのidf_{all}をidf_{genre}で割った値である(数式[1]参照)．このとき，全テキストでの出現テキスト数が低く，特定ジャンルの出現テキスト数が高い語彙の数値が高くなる.ただし，出現語彙数自体が少なくキーワード性の低い場合にもidf値は非常に大きくなるため，各ジャンルテキスト群中での出現テキスト比率が閾値を超える単語のみをジャンルの特徴的語彙として扱う．本研究においては，閾値は10%とする．

$$CIDF=\frac{idf_{all}}{idf_{genre}}=\frac{\log\left(\frac{D_{all}}{\{d_{all}:t_i\in d_{all}\}}\right)}{\log\left(\frac{D_{genre}}{\{d_{genre}:t_i\in d_{genrel}\}}\right)} \quad [1]$$

各ジャンルの名詞でCIDFの上位10語を，全体とそのジャンルのみでの各出現頻度とそれぞれのidf値と合わせて表2から表8に示す．

ネットワーク図の場合と異なり，CIDFに基づいたジャンルに特徴的なキーワードには，批評テキスト全体に共通する単語が少なく，代わりにほとんど個々のジャンルのみに出

現する単語の比率が高くなっていることがわかる.

表2　エンターテイメントの特徴的語彙

	出現頻度		idf		
	all	genre	idf_{all}	idf_{genre}	CIDF
小説	4117	1701	0.92	0.67	1.38
物語	2300	919	1.31	1.06	1.24
短編	738	293	2.43	2.07	1.17
長編	749	285	2.20	1.91	1.15
作者	934	374	2.02	1.75	1.15
主人公	1582	563	1.60	1.40	1.14
男	1330	501	1.87	1.67	1.12
女	784	301	2.33	2.12	1.10
自分	967	330	2.06	1.88	1.09
今月	678	235	2.14	1.98	1.08

表3　歴史の特徴的語彙

	出現頻度		idf		
	all	genre	idf_{all}	idf_{genre}	CIDF
時代	1018	172	2.00	0.70	2.84
江戸	135	47	4.06	1.69	2.41
小説	4117	277	0.92	0.43	2.12
ラスト	255	37	3.13	1.91	1.64
次	326	45	3.02	1.88	1.61
連作	601	74	3.00	1.91	1.57
長編	749	75	2.20	1.43	1.54
子	323	34	3.24	2.11	1.54
造形	434	52	3.31	2.18	1.52
新人	337	41	3.06	2.07	1.48

表4　純文学の特徴的語彙

	出現頻度		idf		
	all	genre	idf_{all}	idf_{genre}	CIDF
作品	2886	467	1.21	0.62	1.95
文学	563	145	2.62	1.44	1.82
物語	2300	371	1.31	0.73	1.81
小説	4117	509	0.92	0.55	1.66
自分	967	152	2.06	1.41	1.46
作家	1930	235	1.46	1.03	1.42
愛	336	68	3.08	2.17	1.42
現代	624	102	2.35	1.68	1.40
世紀	379	59	2.92	2.13	1.37
世界	1329	193	1.75	1.28	1.37

表5　恋愛の特徴的語彙

	出現頻度		idf		
	all	genre	idf_{all}	idf_{genre}	CIDF
恋愛	754	277	2.99	0.53	5.66
恋	296	78	3.24	1.10	2.95
小説	4117	234	0.92	0.41	2.27
愛	336	55	3.08	1.39	2.22
恋人	208	30	3.47	1.79	1.94
女	784	72	2.33	1.21	1.93
ヒロイン	466	52	2.91	1.57	1.85
男	1330	103	1.87	1.06	1.77
女性	718	52	2.36	1.50	1.57
中年	195	18	3.51	2.26	1.55

表6　ミステリーの特徴的語彙

	出現頻度		idf		
	all	genre	idf_{all}	idf_{genre}	CIDF
ミステリー	1489	1038	1.83	0.71	2.57
事件	1170	701	2.03	1.03	1.98
作品	2886	1168	1.21	0.68	1.79
殺人	519	343	2.63	1.52	1.73
探偵	813	592	3.01	1.76	1.71
捜査	552	459	3.37	2.07	1.63
読者	1186	524	1.82	1.16	1.57
主人公	1582	627	1.60	1.03	1.55
シリーズ	990	533	2.21	1.44	1.53
謎	697	341	2.34	1.56	1.50

表7　SFの特徴的語彙

	出現頻度		idf		
	all	genre	idf_{all}	idf_{genre}	CIDF
SF	2040	1403	1.84	0.27	6.71
宇宙	310	241	3.30	1.61	2.05
未来	291	159	3.18	1.77	1.80
ネタ	326	171	3.02	1.69	1.78
星	216	139	3.51	1.97	1.78
地球	145	97	3.85	2.27	1.69
設定	849	285	2.69	1.73	1.55
文庫	412	148	2.77	1.79	1.54
日本	1011	221	2.04	1.50	1.36
時間	322	96	3.06	2.28	1.34

表8　ノンフィクションの特徴的語彙

	出現頻度		idf		
	all	genre	idf_{all}	idf_{genre}	CIDF
本	1657	867	1.61	0.88	1.82
エッセイ	224	165	3.39	2.21	1.53
ノンフィクション	179	133	3.58	2.39	1.50
著者	827	397	2.12	1.42	1.49
人	1702	680	1.52	1.07	1.42
写真	196	139	3.79	2.76	1.37
取材	244	160	4.03	2.98	1.35
対談	99	85	5.03	3.73	1.35
日本人	219	123	3.53	2.65	1.33
インタビュー	135	95	4.51	3.39	1.33

一方で，エンターテイメントと純文学のジャンルでは，各ジャンルに特徴的な語彙よりも全体にわたって特徴的な語彙の方が多く見られる．これらのジャンルは一つのジャンルとして認識されているとはいえ，その内容は多岐に渡る．このため，これらのジャンルを批評する場合にも，固定的な切り口で語ることが困難であり，結果として，ジャンルに特徴的な語彙が抽出されにくいのであると考えられる．

ジャンルに特徴的な語彙が示す概念は，各ジャンルを批評する上で，中心となる概念であると考えられる．これらの語彙に着目し，そのテキスト内での使われ方を計量的に分析することで，各ジャンルにおける読解と批評のメカニズムの差異を明らかにすることが可能であると考えられる．

5　結論と今後の課題

本研究では，複数ジャンルに渡る書評テキストを横断的に計量分析することで，各テキストジャンルの批評における中心的なテーマの異なりと，それらの関係性をネットワーク図として表現した．

また，各ジャンルの批評において特徴的な語彙を抽出するため，IDFよりCIDFを指標として作成し，CIDFによって特徴的語彙を抽出した．

今後は，これらのジャンルに特徴的な構造と語彙を用いて，テキスト批評での認知的なメカニズムを計量的に明らかにすることを目指す．

謝辞

本研究は科研費「知識共有のための価値指向型オントロジーの多分野多言語化」(20300074) の助成を受けた．

参考文献

[1]James Rye, Peter Rubba: “Scoring Concepts Maps: An Expert Map-Based Scheme Weighted for Relationships”, School Science and Mathematics, pp. 33-44, 2002.

[2] 佐藤さやか，松本斉子，村井源，徃住彰文: ″テキストの内容分析による文芸批評の認知構造の推定″, 第10回日本感性工学会大会論文集, P01-11, 2008.

[3] 本の雑誌: http://www.webdoku.jp/mettakuta/, (2009 年1 月30 日参照).

[4] 村井源，松本斉子，徃住彰文: “コーパスに基づく感性と論理のネットワーク表現”, 日本認知科学会第25回大会発表論文集, 194-197, 2008.

村上春樹の初期三部作における構造解析

Structual Analysis for an Early Trilogy by Haruki Murakami

工藤彰[1*], 村井源[1], 徃住彰文[1]
Akira KUDO, Hajime MURAI, Akifumi TOKOSUMI

*1 東京工業大学大学院社会理工学研究科
Graduate School of Decision Science and Technology, Tokyo Institute of Technology
〒152-8552 東京都目黒区大岡山2-21-1
E-mail: {akudo,h_murai,akt}@valdes.titech.ac.jp

本研究では村上春樹の初期三部作を用いて,語彙の出現頻度や語彙同士の関係性に着目し,単語の計量分析からネットワークを作成した.主として,三部作に共通して登場する人物である「鼠」がネットワークの中心を占めていることに着目し,共起単語より「鼠」の役割を分析した.その結果,三部作中の第二作途中をさかいに,物語の中心に位置していた「鼠」の出現頻度が落ち,中心から周辺に移行していく構造が見出された.また,従来の文学批評でなされてきた言説と計量的分析を比較することで,本研究の文学的解釈における有効性が確認された.

This study analyzes an early trilogy by Haruki Murakami using measurements for word occurrences and co-occurrences, and constructs word networks for the three novels. We particularly focus on the common character Nezumi throughout the three novels, and analyze the co-occurred words for Nezumi. The results indicate that the frequency with which the name Nezumi appears corresponds strongly to the centrality or peripherality of his role within the novel. Finally, through a systematic comparison of our quantitative methodology to more traditional methodologies within literary studies, we discuss the effectiveness of our approach for literary interpretation.

キーワード: 村上春樹, 文学, 文芸批評, テキスト解析, 計量書誌学
Haruki Murakami, literature, literary critique, text analysis, bibliometrics

1 はじめに

日本の現代文学を語る上で村上春樹の名前をいまや無視することは不可能である.国内での谷崎潤一郎賞や読売文学賞の受賞をはじめ,フランツ・カフカ賞やフランク・オコナー国際短篇賞,そして記憶に新しい2009年初頭のエルサレム賞の受賞等,近年は特に海外での評価と知名度が急激に上昇している.また,彼の小説に影響を受けたと言われる作家も増加しており,円熟期を迎えた村上春樹とその小

説に対する関心は高まる一方である.そこで,いま,再び村上春樹という作家のテキストに新たな分析のスポットライトを当て,これまで試みられていなかった解釈の地平を探ることには十分意義があるように思われる.

2 対象テキスト

本研究で対象としたのは,村上春樹の長篇小説の中でも,一般的に初期三部作と呼ばれる『風の歌を聴け』[1]『1973年のピンボール』[2]『羊をめぐる冒険』[3][4]の三作品である.これらの作品には文芸批評家や文学研究者のあいだでも,すでに数多くの論究がなされており,コンピュータを用いた文学研究との差異を明確に打ち出すには極めて適当な対象テキストである.

さらに本研究が,方法論先行の視野の狭い論考に陥るのを避けるため,長い歴史と伝統を持つ文芸批評との対比を試みた.テキストとして用いたのは『ハルキ・ムラカミと言葉の音楽』[5]『村上春樹「喪失」の物語から「転換」の物語へ』[6]『村上春樹 イエローページ①』[7]の三冊である.

3 計量的方法

3.1 共起頻度

具体的な方法として,どのような言葉を用いて作家が小説を構築しているのかを分析するために,作品の文章中全体から計量的に語句を抽出した.今回は,文章を形態素に分け,さらに共起語のつながりを発見するため,CaboCha[8]を使用した.

また本研究では主に重要語句を見出しやすい名詞を利用対象とした.このような計量的手法を扱った先行研究としては,村井らが異本テキストをもとにした思想的メッセージの抽出を行っている[9].

まずは,初期三部作中で主人公同様の扱いを受ける,きわめて重要な登場人物の「鼠」に着目し, 共起頻度を測った.以下の表は,各作品における「鼠」の共起語の上位約十語である.なお,『風の歌を聴け』は「風の歌」,『1973年のピンボール』は「1973」,『羊をめぐる冒険』は「羊」と表記した.

表1 初期三部作中の「「鼠」の共起語

風の歌		1973		羊	
ビール	11	ジェイ	15	羊	12
首	6	ビール	15	ビール	10
家	6	グラス	14	手紙	9
バー	5	煙草	13	男	7
カウンター	5	目	9	部屋	7
口	4	女	8	山	7
本	4	ピンボール	8	本	5
父親	4	車	8	音	5
車	4	心	8	前	5

三作品に共通して頻出するのは,「ビール」「煙草」といった語である.これは「鼠」がバーで飲んでいる場面が,作中で頻繁に見られることに関連していると考えられる.

3.1.1 『風の歌を聴け』

まず『風の歌を聴け』に注目すると,「首」という語が上位に目立つ.これは主に文中で「首を振る」という表現で現れている.物語における主要な登場人物である「鼠」に「首を振る」行為をさせるのは,自分の周囲の人々や環境,世界に対し否定的な態度をとることを意味する.

それはまた,実家が金持ちであるという「鼠」が己の出自を否定するという物語の内容に一致する.くわえて,そのことは「首」という語だけから読み取れるわけではない.たとえば,好んで堕落した生活に身を浸しているという点で,嫌悪の対象である経済的に豊かな自分の「家」や「父親」という語に共起していることも多いに関係しているだろう.

3.1.2 『1973年のピンボール』

また『1973年のピンボール』において共起頻度の高い「女」は,「鼠」と短期間ではあるが付き合う女性のことである.「ピンボール」はタイトルにも含まれているとおり,重要なキーワードとして物語も後半に入ってから,主人公が不意に探し求めることになる事物のことである.

3.1.3 『羊をめぐる冒険』

最後に『羊をめぐる冒険』には「羊」「手紙」「男」「部屋」「山」等が際立って多くあらわれる.特に「羊」という語に関しては,やはり題名の一部になっているが,『1973年のピンボール』の「ピンボール」同様,主人公がストーリー上,追い求めることになる対象である.

以上のことからわかるよう,その物語に固有の特徴を示すキーワードが,重要な登場人物「鼠」の共起語の中から,十分計量的に掬いとることができたと言える.

3.2 共起ネットワーク

次に,計量した語彙の出現頻度と共起頻度をもとに,ネットワーク図を作成した.ノードの大きさは語彙の出現頻度に近似するようにレイアウトされる.また,エッジは共起している語同士をつないだものであり,その語と語の近さは共起頻度に対応している.なお,出現頻度は多いが共起のない語は図から除外した.

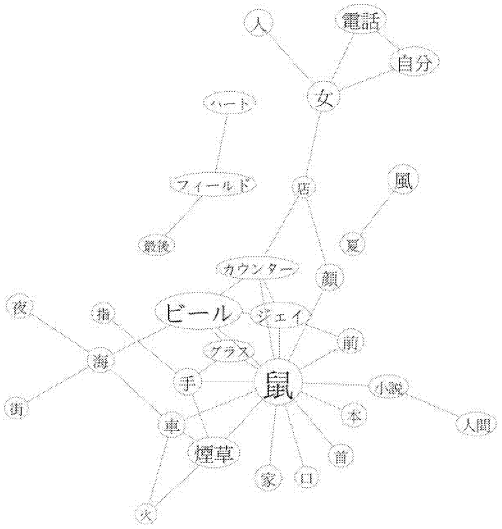


図1　『風の歌を聴け』の共起ネットワーク

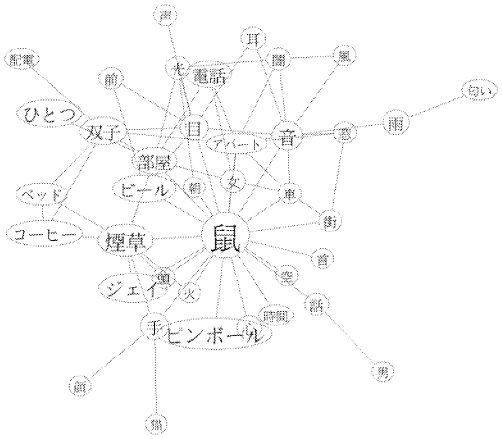


図2　『1973年のピンボール』の共起ネットワーク

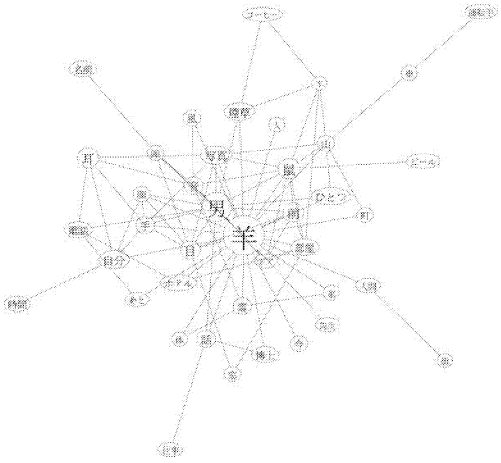


図3　『羊をめぐる冒険』の共起ネットワーク

それぞれのネットワーク図を比較すると，図1と図2において中心語であった「鼠」が，図3では中心からやや離れた位置に移行しているのが見てとれる．

さらに，図2の『1973年のピンボール』について，詳細な分析を試みるため，冒頭から12章までを前半，13章から最終章の25章までを後半としてネットワーク図を作成した．

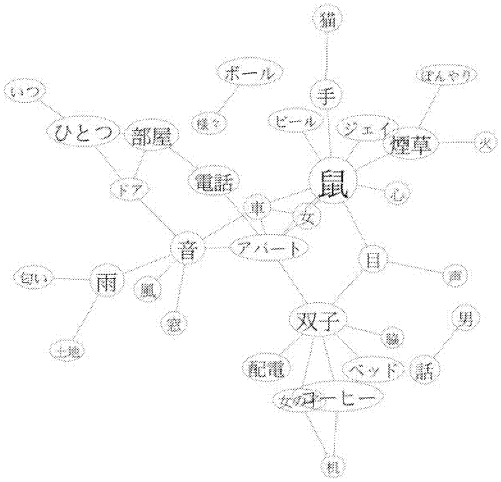


図4　『1973年のピンボール』（前半）の共起ネットワーク

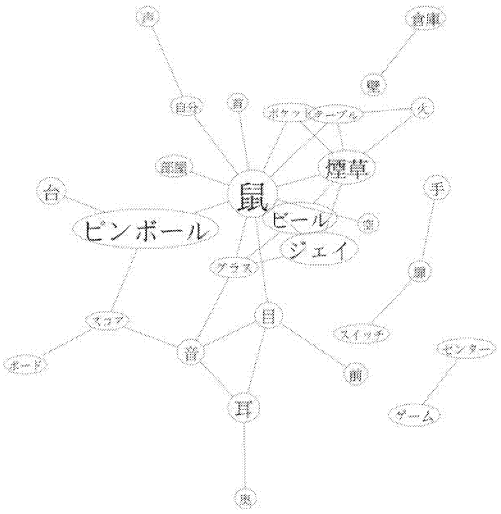


図5　『1973年のピンボール』（後半）の共起ネットワーク

図4と図5を比較すると，図4では中心語「鼠」のまわりには「煙草」「ビール」「ジェイ」等と共起している．しかし，図5になると「ピンボール」という図4では見られなかった新たな語彙と共起しているのがわかる．

この例が示すように一作品中で計量分析を行う場合でも，出現する語彙や共起する語彙がまったく異なる場合があることが確認された．

3.3 考察

初期三部作の内容については，共起ネットワークに注目することで『風の歌を聴け』と『1973年のピンボール』前半までは中心語であった「鼠」が，『1973年のピンボール』後半から『羊をめぐる冒険』にかけて中心から周囲に移行した構図が見てとれた．このことから「鼠」の小説内

での出現頻度は相対的に減少したといえる.

以上のように,各作品ごとの語彙の特徴は,計量的な方法を用い共起語やネットワーク図に注目することで,それぞれの違いを見ることが可能だということが証明された.また,対象とするテキストを必要に応じ部分的に限定していくことで,中心語やネットワーク図も厳密に変化することがわかった.

今回のような計量的方法を用いることで,語彙レベルでの構造の変遷がはっきりとし,作品固有の特徴が明確になったといえる.

4 文芸批評との比較

ここまで論じてきた計量的手法は,客観的な操作を介在させることで,小説内部の構造を語彙レベルから紐解いていけることを証明した.しかし一方で,実際の読み手である人間の理解,それも,より高次な文学解釈を果たして可能にするものなのかどうかはやや未知数にも思える.そこで昨今までの村上春樹に対する言説でどのようなことが語られていたかを検証し,本研究の対象に合わせ初期三部作の批評を取り上げてみたい.

4.1 『風の歌を聴け』批評

Rubinによれば,『風の歌を聴け』には,まわりのものと自分とのあいだの距離を測るためのものさしが重要だというテーマが内在されている.これは「鼠」という人物に対象を与えることで世界の認識を明確にしていくことだと捉えることができる.「鼠」に共起している語からやはり「父親」「家」といった「鼠」の憎むべき対象があり,逆に「ビール」や「煙草」や「本」に満ち足りた生活を愛するという対象が見てとれる.

4.2 『1973年のピンボール』批評

加藤は「鼠」の死を「ピンボール」に置き換えた換喩,代置の物語だと指摘する.「鼠」の死までは単純に語彙のレベルで判断できないものの,「鼠」と「ピンボール」のあいだに強い関連が現れる事実は,本作品を前半と後半にわけた共起語のネットワーク図,図4と図5の比較から容易に認識できる.

4.3 『羊をめぐる冒険』批評

Rubinは,『1973年のピンボール』と決定的に違うのは,幽霊しか住んでいない記憶と死の世界より,普通の肉体が住む現世が好ましいと悟ったところだとまとめる.この指摘の根拠は前二作から,舞台が大幅に転換したことに起因すると思われるが,現実の北海道に向かうという物語の視点にたてば,実際「山」や「羊」などの語がこの作品からはじめて「鼠」に共起している.

4.4 結論

このように文学批評と比較してみると,計量的方法でも十分文学的解釈に迫りうる可能性はあるように思える.当然,語彙の頻度からのアプローチであるゆえ,頻度が少ないにも関わらず物語にとって重要な語句の発見は困難かもしれない.だが,可能な限り数多くの作品を対象とし,テキスト同士の結果を比較検討することで,より個別性が浮き彫りになることは間違いない.

また物語を読解し批評する語彙と,物語に出現する語彙には幾らかのずれがあった.それは,批評が細部だけでなくテキストの全体についても言及を行うからで

ある．より抽象度の高い批評になれば，計量した語彙もそのまま対比するわけにはいかない．計量的方法の対処としては語彙にメタデータを付与し，幾つかの範列的な集合に分け，対立させたり比較したりしていく必要がある．さらに作家自身も小説を豊かなものにするため類義語を戦略的に用いる以上，語彙の意味によるグルーピングは必須になるだろう．

5 今後に向けて

本研究をさらに敷衍させ充実した内容にするために，筆者は今後，対象を村上春樹の長篇全ての11作品にまで拡張し，より全体的なパースペクティブを確保することによって，大規模な分析を計画している．本研究では村上春樹小説の「転換期」と言われる作品群の前後まで扱うことはできなかった．しかし，従来の文学研究者や文芸批評家が村上春樹作品について，通時的な批評や言及をする際，最も議論の焦点になる箇所である以上，計量的手法がどの程度の仔細な発見と検証をもって迫れるかは，非常に興味深いものがある．また，村上春樹に固有の特徴をより客観的に示すためにも，同時代の作家のテキストを比較対象にするといった可能性も視野に入れていきたい．

謝辞

本研究は科研費「知識共有のための価値指向型オントロジーの多分野多言語化」（20300074）の助成を受けた．

参考文献

[1] 村上春樹：『風の歌を聴け』，講談社文庫，160p.，2004.
[2] 村上春樹：『1973年のピンボール』，講談社文庫，183p.，2004.
[3] 村上春樹:『羊をめぐる冒険（上）』，講談社文庫，268p.，2004.
[4] 村上春樹:『羊をめぐる冒険(下)』，講談社文庫，257p.，2004.
[5] Jay, Rubin（畔柳和代訳）：『ハルキ・ムラカミと言葉の音楽』，新潮社，463p.，2006.
[6] 黒古一夫：『村上春樹「喪失」の物語から「転換」の物語へ』，勉誠出版，293p.，2007.
[7] 加藤典洋：『村上春樹 イエローページ①』，幻冬舎，249p.，2006.
[8] Cabocha/南瓜:Yet Another Japanese Dependency Structure Analyzer, http://chasen.org/~taku/software/cabocha/（2009 年4 月1 日参照）
[9] 村井源，徃住彰文：「正典テキスト群から編集的中心メッセージを抽出するネットワーク解析法」，情報知識学会誌，Vol. 17, No. 3, pp.149-163, 2007.

文学作品群の特徴的語彙と概念カテゴリーの抽出
― 星新一ショートショートの計量分析―

Extracting Distinctive Vocabularies and Conceptual Categories from a Literary Text: A Quantitative Analysis of Hoshi Shinichi's Short-Short Stories.

佐藤知恵*[1]、村井源、徂住彰文
Chie SATO, Hajime MURAI, Akifumi TOKOSUMI

1　東京工業大学大学院社会理工学研究科
Department of Value and Decision Science, Tokyo Institute of Technology
〒152-8552 東京都目黒区大岡山 2-12-1-W9
E-mail: {chiesato, h_murai. Akt}@valdes.titech.ac.jp

本研究では，特定の作家の小説を特徴づける語彙とそれが所属する概念カテゴリーの特徴を調べるため，星新一ショートショート作品の特徴的概念と，その構造を計量的に抽出した．データとしては，ショートショートの第一人者である星新一氏の初期（1968 年～1973 年）における全 720 作品を電子化したテキストを対象とし，出現頻度の高い語彙の共起ネットワーク化を行うことで，星作品における概念構造を可視化した．また，各作品に特徴的な語彙を抽出し，そのカテゴリー分類を行うことで，ショートショートの特徴を計量的に抽出した．

The purpose of this research is to identify the distinctive words and concepts that characterize a literary style of an author. The present study analyzes Shinich Hoshi's early 720 short-short stories in three methods; (i) Based on co-occurrence frequencies, the main vocabularies and their relationships are visualized as a network, (ii) Employing the TF/IDF measure, a set of words distinctive to a story is extracted, (iii) Conceptual categories of each distinctive word are analyzed according to a thesaurus definition of category levels. The paper argues constraint of the length of stories and its effects on the selection of words by the author.

キーワード：星新一，ショートショート，文学，ネットワーク，共起単語
Shinichi HOSHI, Short-short, literature, network, co-occurrence

1 はじめに

ショートショートの第一人者である星新一の作品は,出版から年月を経た現在でも世代を超えたファンが多く,作品が実写化されるなど,話題に事欠かない.星作品は,SFものの多さや,淡白な表現手法,オチがあることなどが特徴とされ,批評などで数多く言及されてきた.しかし,その特徴が,どの程度テキスト中に表出し,それぞれがどの程度の影響力をもつかについての詳細については未だ解明されていない点が多い. 本研究では,星作品の語彙と概念カテゴリーの特徴を計量的に調べ, 特徴的キーワードの抽出と,概念カテゴリーの可視化によって,星ショートショートがもつ特徴を計量的に分析する. またその考察から物語分類の基盤を示す.

星新一ショートショートを対象とする意義としては,以下が挙げられる.

1)星新一氏は日本におけるショートショートの開拓者・第一人者であり, ショートショートの典型的データと言える.

2)星氏のショートショート作品は全 1280 編, 文字数約 237000 が残されており, 計量的な解析にあたり, 信頼できるデータ量が確保できる.

3)ショートショートの文体の特徴として, 明快で淡白な表現が多く, 複雑な修辞的テクニックが抑制される傾向があり[1], 計量的解析との相性が高い.

2 対象データ

星新一のショートショート作品全1280編[2]のうち, 本研究では, 星氏の全執筆活動期間(1961～1997)を前期(1961～1973)と後期(1974～1997)に分け, 作品執筆率の高い前期の 720 作品を電子化し, 対象とした. 星氏の作品の年代別短編発表数を図1に示す.

また, テキスト化されたデータから, 各作品の文字数のばらつきを調べた(図 2). その結果, 最も多くの作品数が属するデータ区間は, 2400～2800(原稿用紙 6～7 枚分)であり, 文字数の平均値は 4607(原

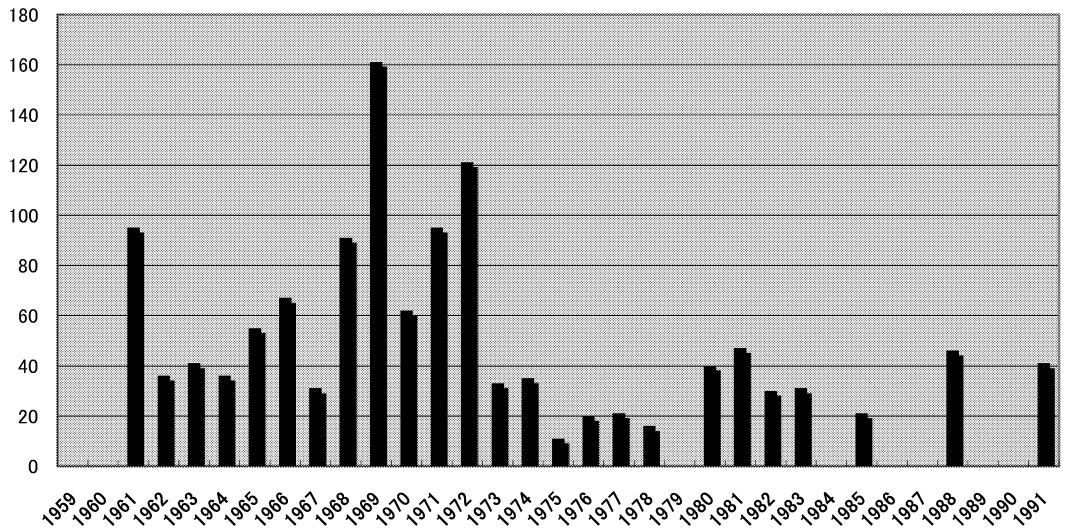


図 1 年代別ショートショート作品発表数(作品数/発表年)

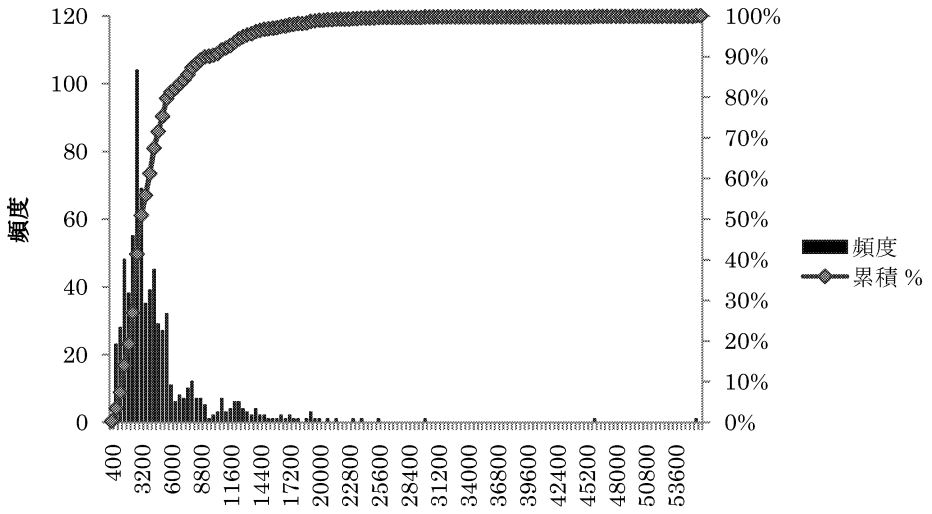


図 2 文字数別作品頻度(作品数/文字数区間)

稿用紙 6 枚弱)であった. さらに, 図 2 から分かるように, 星氏のショートショート作品は, 1600～5600 までの文字数のものが大半を占めており, 原稿用紙 4～14 枚が星新一ショートショート作品の代表的な“長さ”だと言える.

3 分析方法

本研究で用いた分析方法は, 頻出語彙ネットワーク分析, TF/IDF, カテゴリー分類である.

3.1 ネットワーク分析

ネットワーク分析では, 内容的概念構造を計量的に分析するため, ネットワーク状に可視化する[3]. この手法を用いて, 星作品における共起頻度の高い上位 50 の名詞の関係性を図 3 に示した.

図 3 においては, 頻出語彙の影響力を可視化するため, ネットワーク図でのノードのフォントサイズは, 単語の出現頻度の二乗根に比例させている. このネットワーク図において, 星氏の作中人物として, 男, 女, 青年など抽象的な人物描写を選択的に設定する傾向にあることが分かる.

しかし, その一方, 博士, 装置など, 特徴的なキーワードへの偏好も見られた.

3.2 TF/IDF

ここで, TF/IDF を用いて, 各作品に特有の語彙(他の作品には出ていないが, その作品にはよく出ている語彙)を抽出し, その結果を, 出現頻度に基づく分析結果と比較した. 表 1 は、二つの分析において値の高い上位 20 の名詞を比較したものである. 出現頻度による結果では, 顔, 頭, 口など, 一般的に出現頻度が高くなりがちな名詞が上位に入っていると考えられる. TF/IDF による結果では, これらの語彙が抜け落ち,「博士」,「星」といった, 星新一の物語世界の特徴を示す語彙群が高い値を示していることが分かる. 表中の TF/IDF 値は, 上位 10 語を各作品でとった場合の頻度である, 単純頻度では上位に出なかったが, TF/IDF では上位に入った名詞は, 表中の太字で示された 5 項目である.

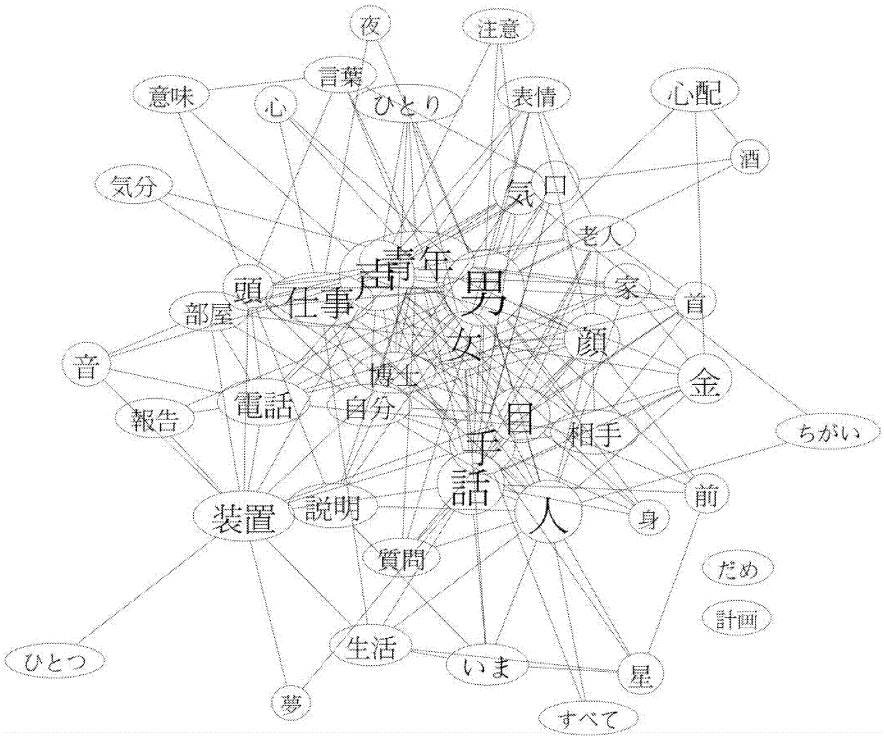


図 3 星作品の語彙ネットワーク

3.3 カテゴリー分類

さらに,より詳細な特徴を追跡するため, この 5 名詞について, 分類語彙表[4]に基づくカテゴリー分類を行い, 考察した[5]. また, 19 位の老人については, 属する部門(人間)の名詞が上位 20 項目中 5 項目(青年,女など)を占めることから注目し, 詳細に分析を行った

分類語彙表のカテゴリー構成は, 類, 部門, 中項目, 分類項目の四つから成り, この順にカテゴリーの詳細度は増す. 例えば, 警察官という名詞は分類番号 1.241701 で分類される. 左から, 1 は類(体の類), 2 は部門(人間活動の主体), 4 は中項目(成員), 17 は分類項目(保安サービス)で表現され, 01 により警察官, 保安官などと同一のカテゴリーに所属していることを示す. 本研究では, これら 4 段階のうち, 中項目と部門を利用する.「博士」「星」「夢」「老人」「家」の 5 名詞が属するカテゴリーはそれぞれ,「人物(1.2340)」,「天体(1.5210)」,「飢渇・酔い・疲労・睡

表 1 上位語彙の比較

順位	出現頻度	TF/IDF
1	男(1843)	男(140)
2	声(1575)	声(130)
3	人(1228)	人(92)
4	話(1203)	話(87)
5	手(1183)	手(74)
6	女(1036)	女(73)
7	青年(1009)	青年(73)
8	目(950)	装置(58)
9	仕事(881)	仕事(56)
10	顔(771)	**博士(53)**
11	装置(766)	**星(51)**
12	金 (779)	電話(50)
13	気(701)	金(41)
14	電話(649)	目(40)
15	頭(631)	相手(32)
16	説明(570)	顔(30)
17	音(563)	音(28)
18	相手(561)	**夢(25)**
19	心配(552)	**老人(25)**
20	口(542)	**家(24)**

眠など(1.3003)」「老少(1.2050)」,「家(1.4400)」であり,老人が属する部門は,「人間」であった(カッコ内は分類番号).これらのカテゴリーについて頻度上位名詞を調べ,この結果を表 2～7 に示した.表 2～7 中の頻度については,全 720 作品での出現頻度を示す.

4 考察

4.1 人物カテゴリー

人物カテゴリーでは,「博士」や「先生」,「学者」など,人物の職業の頻度が高い.登場人物設定を行う際,職業による人物設定が最も端的かつ効果的に表現できるためだと考えられる.

人間性に深入りしない,職業による抽象的な描写選択には,短いショートショートの中に最大限のストーリー展開を実現するための手法が表れていると考えられる.

さらに,「強盗」,「泥棒」,「囚人」など,犯罪性・事件性の高い人物も多い.星氏の作風として日常から非日常へ物語を急展開させるため,こういった人物の出現頻度が高くなると考えられる.

4.2 天体カテゴリー

天体カテゴリーでは,星,地球など,宇宙的視点の多さを意味し,このカテゴリーの頻度の高さは,星作品の SF 的特徴を計量的に裏付けている.

4.3 飢渇・酔い・疲労・睡眠カテゴリー

飢渇・酔い・疲労・睡眠カテゴリーでは,夢の頻度が著しく高い.このカテゴリー頻度の高さは,星作品のファンタジー的特徴を裏付けている.

4.4 家カテゴリー

家カテゴリーでは,家に続いて城,屋敷といった古い時代を連想させる語彙が続く.

表 2 人物カテゴリー　表 3 天体カテゴリー

語彙	頻度
博士	489
先生	154
患者	125
強盗	108
学者	91
ばか	87
惑星	79
泥棒	71
虫	36
囚人	35

語彙	頻度
星	491
地球	275
太陽	35
満月	7
火星	6
衛星	5
星座	5
星	4
恒星	4
小惑星	3

表 4 飢渇カテゴリー　表 5 家カテゴリー

語彙	頻度
夢	337
退屈	41
眠り	41
麻酔	36
退屈	32
冬眠	32
悪夢	30
疲れ	29
疲労	28
酔い	25

語彙	頻度
城	145
屋敷	62
アパート	53
自宅	41
住宅	30
別荘	26
住居	24
マンション	24
宅	11
巣	10

これは,星ショートショート作品集で,江戸に時代設定されたものがあり,ここから大きく影響を受けたと考えられる.また,「家」の出現頻度が高いことは,空間の限定を意味しており,ショートショートにおいてストーリーの過剰な広がりを内容的・空間的に抑える役割も果たしていると考えられる[1].星氏の作品には,家の中や『ドアのこちら側』など,空間設定されたものが多く,これも,ショートショートの特徴と言える.

4.5 老少カテゴリーと人間部門

老少カテゴリーでは,顕著な特徴は見られないが,老少カテゴリーの上位カテゴリーである,人間部門でみた場合,神,悪魔,妖精,鬼など,超人的な人物選択を確認できる.人間部門において,ファンタジー的特徴が顕著に現れていることが分かる.

表6　老少カテゴリー

語彙	頻度
老人	371
子供	205
女の子	121
少年	106
坊や	94
老	75
若者	52
娘	40
死者	34
少女	22

表7　人間部門

語彙	頻度
男	1843
女	1036
自分	519
老人(老)	446
子供	205
女性	188
神	132
女の子	121
少年	106
坊や	94
悪魔	93
美人	85
妖精	82
鬼	77
他人	69

5　結論と今後の課題

本研究では，星作品を計量分析することで,ショートショートの特徴的な語彙とそれらの関係性をネットワーク図として表現した.また考察から,星ショートショートの特徴として,以下のことが結論づけられた.

1. 長さは文字数 2400～2800(原稿用紙 6 枚～7 枚)のものが一般的である.

2. 短い物語の中に最大限のストーリー展開を埋め込むための,語彙と概念カテゴリーの特徴的選択が計量的に確認できる.

3. 彼の作品テーマとして主要とされる「SF 的特徴」と「ファンタジー的特徴」が、TF/IDF による抽出語彙のカテゴリー分類によって，判別し得る.

4. 星氏が起用する作品テーマとして,SF やファンタジー以外に,事件性が挙げられる.

この結果をもとに，今後は，日常から非日常を誘発する SF，ファンタジー，事件の三点を軸とした分析を進めるとともに，これを基盤とした分類方法を模索し，このジャンルに特徴的な物語構造のパターンの抽出を目指す.

謝辞

本研究は科研費「知識共有のための価値指向型オントロジーの多分野多言語化」(20300074)の助成を受けた.

参考文献

[1]長田貴誉子:「星新一の作品分類と分析試論　一〇〇一編ショート・ショート群を考える」，国文学会誌，Vol.22，pp.76-115，1993.

[2] 星新一:「星新一ショートショート 1001」，新潮社，1998

[3] 村井源;松本斉子;徃住彰文:「コーパスに基づく感性と論理のネットワーク表現」，日本認知科学会第 25 回大会発表論文集，pp. 194-197，2008

[4] 国立国語研究所:「分類語彙表」,『国立国語研究所資料集』 14，国立国語研究所，2004

[5] 村井源;松本斉子;山本竜大;徃住彰文:「Web の計量言語学的分析からみた政治的感性の特徴」，感性工学会研究論文集，Vol.7，No.3，pp. 561-569，2008

音楽評論論文にみる概念構造の変遷
—ネットワーク中心性を用いた音楽概念の抽出—

Conceptual Transitions in Papers of Music Criticism: Extracting Musical Concepts based on Network Centrality

河瀬彰宏[1*]，村井源[1]，徃住彰文[1]
Akihiro KAWASE, Hajime MURAI, Akifumi TOKOSUMI

*1 東京工業大学大学院社会理工学研究科
Department of Value and Decision Science, Tokyo Institute of Technology
〒152-8552 東京都目黒区大岡山2-12-1
E-mail: {kawase, h_murai, akt}@valdes.titech.ac.jp

音楽の様式は刻一刻と連続的な変化を遂げていくため，これらに対して，音楽美学や音楽史学において議論されるような区分，例えば，古典主義音楽，ロマン主義音楽といった確定的な定義を与えることは難しい問題である．しかしながら，時代によって特徴や差異が確認されることは否めない．本研究では，音楽概念を構造的に抽出・体系化することで，音楽概念がどのような差異・変遷をもつものか検討する．具体的には，音楽評論雑誌『ポリフォーン　音楽評論の開かれた場』全 13 冊から，言及対象が 20 世紀以前の西洋音楽に該当する記事と，20 世紀以後の西洋音楽に該当する記事とに試験的に分類し，ネットワーク分析を用いて語彙間の共起関係からそれぞれの音楽評論の思考の中心概念を構造的に抽出する．結果，前者は「世紀」「時間」，後者は「音」「主義」を音楽概念の中心性に強く据えていることが確認された．

This study analyzes the complete thirteen volume set of “POLYPHONE”, which is a highly respected journal of music criticism within Japan, in order to illustrate and highlight differences between and transitions across periods. To that aim, the paper tentatively categorizes articles into two groups, distinguishing between those referring to pre-twentieth century classical music and post-twentieth century classical music. A network analysis is applied to extract the structure of central concepts within the thoughts of musical critics from the texts of journal articles. The analysis results reveal that the primary key word is ‘music’; with ‘musical composition’, ‘creation’ and ‘period’ being strong adjacent words for both concepts. Moreover, while ‘century’ and ‘time’ conceptualize the thinking of the earlier period, ‘sound’ and ‘-ism (principle)’ conceptualize the later period.

キーワード: 西洋音楽，音楽評論，ポリフォーン，概念，ネットワーク分析
Classical music, Music criticism, POLYPHONE, Concept, Network analysis

1 はじめに

音楽の様式・思想・哲学は刻一刻と連続的な変化を遂げていくため，これらに対して，音楽美学や音楽史学において議論されるような区分，例えば，古典主義音楽，ロマン主義音楽といった確定的な定義を与えることは難しい問題である．

しかしながら，時代によって特徴や差異が確認されることは否めない．そこで，音楽史学が従来とってきた，いわばtop-down型の研究方式に対して，本研究では，実データに基づき，計量的に音楽概念を抽出し，構造的に体系化することを試みる．そして音楽概念がどのような差異・変遷をもつものか検討する．

人間の高次感性と高度な知識操作を伴う知的行為のひとつ音楽鑑賞では，聴覚（場合によっては視覚）を動因した知覚体験が必要である．これに加え，楽曲の歴史的背景，作曲家の思想，演奏者の哲学などを汲み取った音楽作品の内容理解も必要となり，音楽体験から，高次な思考プロセスを経た人間の感情喚起を観察することができる．認知科学の立場では，人間特有の高度な知識処理や感性・感情変化の過程をテキストから抽出するが，本研究ではその方法論に則る．

今回，音楽に関する知識を多分に含んだプロの音楽評論家によって語られたテキストを収集し，言及対象が20世紀以前の音楽と，20世紀以後の音楽とに試験的に分類した．

ここに計量的な分析，自然言語処理技術によって単語の出現頻度抽出およびネットワーク分析を行うことで，語彙レベルからそれぞれの概念構造を整理・記述する．

2 テキストの計量分析

テキストの計量分析とは，テキストデータに対して情報理論や統計学を適用し，計量的な解析を行うことで，その特徴を抽出する分析方法である．

質・量ともに豊富なテキストデータから特徴を抽出する主な方法は，テキストに対して構文解析，形態素解析，係り受け解析などを行うことによって，文と文の関係，品詞の集計，単語の出現頻度，単語同士の共起頻度，修飾語被修飾語の関係を計量することである．

例えば，Murai and Tokosumiは聖書と教義文書のテキストの引用関係をネットワーク分析によって計量的に解析し客観的な評価を与えている［1］．

3 本研究で扱う対象

3.1 テキストコーパスの構築

本研究では，1987年10月から1993年12月号までの期間，TBSブリタニカから発行された音楽評論雑誌『ポリフォーン　音楽評論の開かれた場』全13冊を分析対象とし，テキストコーパスを構築する．数ある音楽評論形態の中から雑誌『ポリフォーン』を分析対象として採用した理由は以下の二点に集約することができる．一点目に，近藤譲，柴田南雄，高階修爾，武満徹，田辺秀樹，林光，門馬直美，吉田秀和など，国内外の著名な評論家が多数執筆に携わっているため，プロの音楽評論家による質の高いテキストが提供されることが挙げられる．二点目に，通常

の評論記事だけでなく，書評，対談，時評，海外論潮（国外の音楽評論に対する批評）など，あらゆる評論形式を網羅しつつ，計量的に分析する上で十分な量のテキストが提供されることが挙げられる．

収集したテキストを，音楽作品・作曲家・様式の考察対象の中心が（A）20世紀以前の西洋音楽に関する記事と，（B）20世紀以後の西洋音楽に関する記事とに分類した．

ただし，考察対象の中心が両方の比較に該当するテキストや，インドの民族音楽をテーマとする（A），（B）いずれにも該当しない記事32篇は今回の対象から除外した．また，翻訳作業の伴った記事や通訳を介した対談記事は，著者以外の手が一段階加えられたテキストであるため，思考概念が直接反映されにくいと考え，扱わなかった．

分析に使用したテキストデータの概要は，（A）評論62編，書評22編，対談3編（合計87編，1.67MB），（B）評論53編，書評16編，対談2編（合計71編，1.38MB）である．

3.2 テキストコーパスの特徴

構築したテキストコーパスの特徴を語彙レベルで示す．品詞の分割には日本語係り受け解析器CaboCha［2］を用いた．出現した名詞の総数は，（A）106,104語，（B）92,256語であった．ただし，今回は元文書の表記を尊重し，例えば「音楽」と「ミュージック」であっても別の語彙として扱っている．

3.3 計量的分析

（A），（B）それぞれについて，名詞の出現回数の上位20位までを表2に示す．上位10位以内の語彙を比較すると，「音楽」が最も多く，「作品」「作曲」「オペラ」「時代」「演奏」といった語彙が共通して多く語られており，（A）で

表1 名詞の言及回数の相違（上位20位まで）

（A）20世紀以前に関する音楽評論				（B）20世紀以後に関する音楽評論			
音楽	2188	人間	527	音楽	3203	世紀	408
世紀	954	音	480	音	982	意味	363
オペラ	767	意味	460	作品	962	芸術	338
時代	701	人	442	作曲	930	ピアノ	314
作品	697	世界	433	オペラ	707	現代	305
演奏	657	主義	395	演奏	617	問題	294
作曲	621	楽器	374	時代	565	人	288
時間	607	ピアノ	368	主義	492	ドビュッシー	275
モーツァルト	602	フランス	361	日本	473	古典	269
ウィーン	533	ドイツ	344	世界	415	ジャズ	247

は「世紀」，「時間」，（B）では「音」「主義」「日本」について他方よりも多く言及されていることが分かる．20世紀以前の音楽を語る上では，モーツァルト，マーラー，J. Sバッハ，ワーグナー，ベートーヴェン，といったいわゆる古典派からロマン派にかけて活躍した作曲家が多く出現し，同様に，20世紀以後の音楽を語る上では，ドビュッシー，シェーンベルク，ブーレーズが多く出現している．

また，（A）では2位だった「世紀」，8位だった「時間」が（B）では11位と27位にあり，反対に（A）では12位だった「音」，16位だった「主義」，49位だった「日本」が（B）では2位，8位，9位にある．他の語彙との関連からどのような位置付けにあるのかを4.2節の中心性分析において確認する．

次に，1文内に共起する名詞の単語対を集計した．上位50組中，20世紀以前の音楽では「音」と共起する語彙が32位に確認されるだけだが，20世紀以後の音楽では3組確認される．次節では，全体の上位約30％の単語対からネットワークを構築し，両者の音楽評論に関する概念構造を抽出することを試みる．

4 抽出されるネットワーク

4.1 ネットワーク化

ネットワーク分析とは，複数のノードとそれらを繋ぐエッジによって構成される構造を記述・分析する方法である．本研究ではテキスト内の出現名詞をノードに，共起関係をエッジとする重み付き無向ネットワークを利用する．したがって，１文内に共起する単語対を，ある概念を説明する際に同時に使われる単語対と解釈するならば，構成されるネットワーク

図1（A）20世紀以前の音楽に関するネットワーク

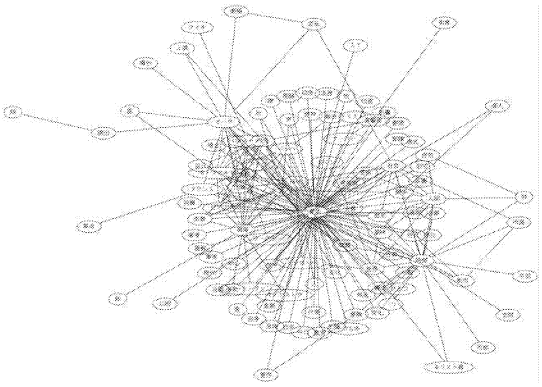
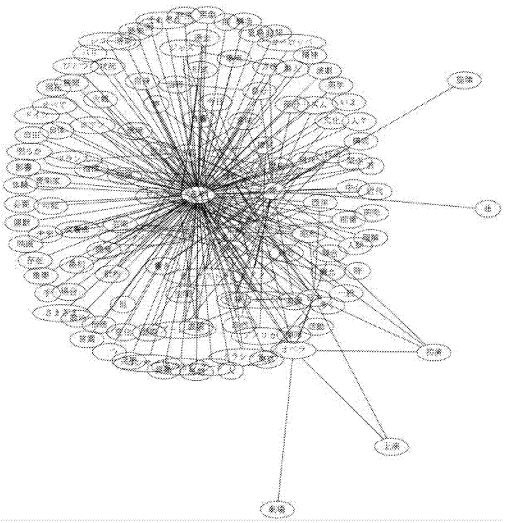


図2（B）20世紀以後の音楽に関するネットワーク



は，評論の執筆者や対談の発言者の思考上に存在する概念同士の関係が表現されるものと解釈される．Graphvizを用いて（A）と（B）の概念関係を視覚化させた図をそれぞれ図1と図2に示す.ただし，ネットワーク図はあくまでも視覚的に把握し易いように，ノードには共起頻度の上位約30%の語彙を，エッジには共起回数が35回以上出現したものを用いている.

ネットワーク図を比較すると，共通して「音楽」を中心に据えて形成されているが，それを直接結ぶ（A）のエッジは密度が低く，対照的に（B）のエッジは密度が高いことが分かる．したがって，20世紀以後に関する音楽評論は，20世紀以前に関する音楽評論よりも，「音楽」と直接的に関わりのある語彙を用いて語られる傾向があると考えられる.

4.2 ネットワークの中心性

中心性分析とは，ネットワーク解析で用いられる手法のうち，ノードの中心概念に基づいて中心の度合いを科学的に測定する方法である．中心概念の捉え方によって中心性の計算方法は多々提案されているが,ここでは代表的な指標として，次数中心性，近接中心性，ボナチッチ中心性［3］の3種類を利用した.

次数中心性はノードがもつエッジの数によって表される中心性，近接中心性はネットワーク全体で他のノードからの平均的な距離が短いものを表す中心性，ボナチッチ中心性は固有ベクトルに基づき，ノードが連結している他のノードの中心性によって決定される中心性である.

これらの指標を評論における思考過程に置き換えると，次数中心性は評論家が言及する語彙量，近接中心性は評論家が近しいと連想する語彙同士の概念，ボナチッチ中心性は，評論家が各語彙の影響を加味した上で語られる概念と解釈することができる.以下の表2にUCINETを用いて算出した各中心性の値を上位10位まで

表2 各中心性とその値（上位10位まで）

	（A）20世紀以前に関する音楽評論			（B）20世紀以後に関する音楽評論		
中心性	次数	近接	ボナチッチ	次数	近接	ボナチッチ
1	音楽 9660	音楽 99.5	音楽 67.1	音楽 14401	音楽 100	音楽 77.5
2	世紀 4395	時代 99.5	世紀 39.5	作品 4876	作曲 100	作曲 40.0
3	時間 3270	作曲 99.0	時代 27.0	作曲 4654	作品 99	作品 39.0
4	作品 3219	世紀 98.5	作品 26.8	音 4208	時代 99	音 31.2
5	オペラ 2934	作品 98.0	作曲 26.5	主義 3016	世界 99	演奏 25.8
6	時代 2870	モーツァルト 97.5	演奏 25.8	時代 2888	現代 97	主義 25.2
7	作曲 2855	意味 97.1	時間 23.6	演奏 2799	音 96	時代 23.8
8	演奏 2679	ドイツ 96.1	オペラ 22.2	オペラ 2728	問題 96	現代 22.1
9	音楽 2387	世界 96.1	モーツァルト 20.6	日本 2611	意味 96	世紀 21.5
10	モーツァルト 2223	演奏 95.2	人間 18.6	世紀 2087	主義 95	日本 21.5

示す.

近接中心性の結果から，（A），（B）は共通して「作品」「作曲」「時代」が強く，ボナチッチ中心性からは，（A）は「世紀」「時間」を，（B）は「音」「主義」を強く中心性に据えていることが確認された.

5 結論と考察

本研究は，音楽評論雑誌を実データとし，計量的に，20世紀以前と以後，2つの時代区分の音楽概念を構造的に抽出した.結果，20世紀以前の音楽を表す概念は「作品」「作曲」「時代」と強く関連し，「時間」「世紀」といった語彙に留意して構築されるものと考えられる．20世紀以後の音楽を表す概念も「作品」「作曲」「時代」と強く関連するが，「音」「主義」といった語彙に留意して構築されるものと考えられる.

武満徹の著作集を対象に，彼の思考の傾向を計量的に抽出したAoshima and Tokosumiの研究[4]から,彼のテキストから抽出される中心概念は「音楽」に次いで「音」が最も多いことが確認されている．20世紀を代表する作曲家であり，かつ，自身の言葉で積極的に音楽を語った武満徹の概念構造も本研究の発見に包括されると言える.

謝辞

本研究は科研費「知識共有のための価値指向型オントロジーの多分野多言語化」（20300074）の助成を受けた.

参考文献

[1] Murai, H. and Tokosumi, A.：“Network Analysis of the Four Gospels and the Catechism of the Catholic Church”, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.11, No.7, pp.772-779, 2007.

[2] Cabocha/南瓜:Yet Another Japanese Dependency Structure Analyzer, http://chasen.org/~taku/software/cabocha/ (2009 年4 月1 日参照)

[3] Bonacich, P.：“Power and Centrality: A family of Measures”, American Journal of Sociology 92, pp.1170-1182, 1987.

[4] Aoshima, Y. and Tokosumi A.：“Extracting Musical Concepts from Written Texts”, The International Conference on Kansei Engineering and Emotion Research 2007 (KEER2007), CD-ROM.

心の状態と言語的特徴 ： ブログにおける商品紹介文の分析

States of mind and stylistic features: The effects of affiliate advertising

斉藤 香里[1*], 村井源[1], 徃住彰文[1]

Kaori SAITO, Hajime MURAI, Akifumi TOKOSUMI

1 東京工業大学
Tokyo Institute of Technology
〒152-8550
東京都目黒区大岡山2-12-1
E-mail: {saitok, h_murai, akt}@valdes.titech.ac.jp

本研究は，自然言語表現から，心的状態が言語表現に与える影響を捉えることを試みる．ブログ記事とアフィリエイト広告を対象として，広告の有無によって，言語表現に異なる特徴が現れるかを検討した．商品紹介記事を分析対象に，計量文体分析と内容分析，修辞表現の比較を行なった．この結果，広告のある記事ほど，記事文中に書き手の受けた印象や雰囲気の記述が多くなった．さらに，書き手の意見や印象を強調する表現が多く現れていた．また，広告のある記事には，逆接を挟んで評価を逆転させる，商品評価の表現が多く見られた．これら表現の差は，ブログ閲覧者の購買行動喚起に対する，ブログの書き手の意欲の差によるものと考えられる．

The purpose of this study is to identify the influence that our states of mind have on voluntary verbalization within everyday situations. The target texts for the present study are blog articles that contain affiliate advertising. The independent variable is whether a blog contains affiliate advertising, while the dependent variables are linguistic features quantitatively measured for the blog articles. Contents analysis, linguistic feature analysis, and rhetoric analysis are carried out on blog articles. These analyses provide the following findings: a) Articles with affiliate advertising have more sensibility oriented expressions. b) Articles with affiliate advertising exhibit greater use of emphasis on the writers' opinions and impression enhancement towards the reader through emphasized expressions. c) Articles with affiliate advertising exhibit more “reversed evaluation expressions”. The paper argues that differences in the intentions of the writers to elicit purchase behaviors in the blog readers produce such differences.

キーワード: 言語的特徴, ブログ, 広告
stylistic features, blog, advertisement

1 目的

心理学実験では，『“ウソ”をつくことに対するモチベーション』の差異によって，言語表現の特徴に差異が生じることが確認されている[1]．しかしながら，この分野の研究はどうしても実験的であり，さまざまな要因の絡み合った実社会において，同様の結果を得られるとは限らない．では，実験統制下にない自然言語表現には，心的状態の差異によって生成される言語内容の特徴に相違が見られるであろうか．この検証は，非常に難しい問題である．そこで本研究では，「心的状況の差異を仮定する指標」を設けることとした．その指標として，インターネット上で公開される個人のブログにおける，商品の紹介・評価を記述した記事と，それに付随する商品広告の存在に注目する．近年では，ブログに商品についての紹介，その使用の評価，感想などを書いて，その商品の販売先にリンクを貼り，ブログの閲覧者がそれをクリック，または経由して利用登録や商品購入を行なうと，ブログの主に金銭的な利益の入る『アフィリエイト広告』が一般的になってきている．[2]アフィリエイト広告の付随したブログにおける商品の紹介・評価記事の文章は，アフィリエイト広告が付随せず，特別な利益の生じないブログでの商品の紹介・評価記事の文章とは異なって，金銭的収入という利益が関わってくる分の心理的バイアスが生じていると仮定できる．すなわち，ブログ記事に付与されるアフィリエイト広告が，広告のない記事との「心的状態の差異を仮定する指標」となる．

本研究では，特定の商品について紹介，または評価を行っている記事を対象として，言語的特徴を分析した．また，アフィリエイト広告を「心的状態の差異を仮定する指標」として捉え，広告の有無による特徴の比較を行なった．本研究は，複数の分析方法，複数の分析視点を置いて分析し，それらを統合して考察することで，可能な限り，言語的な特徴を拾い上げようと試みている．

2 方法

2.1 分析対象

2.1.1 対象商品

特定の商品やジャンルに依存しない，一般的な言語的特徴を得るため，1）一定の知名度があること〔Amazonカスタマーレビューが100件以上付いている，Google検索およびyahoo!検索の結果が500件以上ヒットする〕，2）賛否両論があること〔Amazonの評価が星2つから星4つのあいだにある（但し，全評価階級に投票があること）〕の2つを満たす，3ジャンル×2品目の計6商品を対象とした．

a. 書籍：「女性の品格」／「13歳のハローワーク」
b. TVゲーム：「メタルギア ソリッド 4 －ガンズ・オブ・ザ・パトリオット－」／「ヴァルキリープロファイル2 －シルメリア－」
c. DVD：「ロスト・イン・トランスレーション」／「男たちの大和 －YAMATO－」

2.1.2 広告有無の判断

ブログ記事における広告付与の判断は，次の基準に従って行なった．以下の基準に該当しないブログ記事は，分析対象から除いている．

a.「広告有り」とする判定基準
・ブログ主の手によって付与されたと認められる広告が付いている
・記事内で紹介している商品と一致した，販売先へのリンクがある
b.「広告無し」とする判定基準
・ブログ記事内に，他サイトへのリンクがない

2.1.3 対象記事

ブログ記事中で分析対象となる商品を購入または使用し，その感想を報告しているブログを,分析対象として採取した．対象となるブログ記事はそれぞれ，商品紹介・評価を主目的に公開された，独立した記事である.一商品につき広告有り6記事，広告無し6記事の計12記事，総計72記事を分析対象として採取した．

使用した検索ツールによる結果の偏りを防ぐため，一般的に使用頻度の高いと思われるGoogle検索およびyahoo!検索を使用し，ブログの採取を行なった．

2.3 分析方法

言語的特徴の科学的検討を保証する分析方法として，1）使用される形態素，語彙の分析，2）認知要素の分析，3）記事内容の分析，4）表現方法の分析の4つの方法を選択した．これらの4つの方法は，内容分析の2)・3)と，表現法分析の1)，4)に大別される．本研究では，ブログ記事の内容構成と表現方法に注目し，分析を行なった．

2.3.1 内容分析

2.3.1.1 話題分類

ブログ記事内の話題を分析する単位として，「命題」を認定した．命題分割を行なって得られた，総数1552命題（書籍：478命題，TVゲーム：599命題，DVD：475命題）を分類し，話題構成カテゴリーシステムを作成した（表1）．「対象商品」から「内容関連」までは『「対象商品」に近い話題』，他者評価以下は『「対象商品」から離れた話題』に区分される．

2.3.1.2 認知要素分類

話題分類とは異なる命題分割条件に基づいて，「命題」を認定した．命題分割を行なって得られた，総数2844命題（書籍：889命題，TVゲーム：877命題，DVD：940命題）に対して認知タグを付与し，認

表 1　話題構成カテゴリーシステム

話題	主内容
対象商品	情報
	使用体験
	意見表明
	記述
商品テーマ	情報
	経験
	意見表明
	記述
内容関連	情報
	経験
	意見表明
	記述
他者評価	意見表明
周辺テーマ	情報
	意見表明
	記述
他商品	情報
	意見表明
	記述
その他	

表 2　認知タグ分類表

心的操作主体	大分類	小分類
一人称心的操作	評価	評価判断
		評価感情
	感情・感性	感情
		感性
	思考	思慮
		推測
		推奨
		疑問
		理解
		意欲・願望
		比較
		判断
		個人的見解
	認識	内容記述
		特徴言及
		程度
		属性
		状態・状況
	行為・状態記述	行為・状態
	投影	投影感情
		投影行為
他人称心的操作	他人称感情・感性	他人称感情
	他人称思考	他人称思慮
		他人称推奨
		他人称理解
		他人称意欲・願望
	他人称行為・状態記述	他人称行為・状態
一般的心的操作	一般的行為・状態記述	一般的行為・状態
外部参照知識	外部知識	引用
		情報記述

知タグ分類表を作成した（表2）．

ブログの書き手本人を主体とする認知要素と，それ以外を主体とする認知要素を区別するため，認知活動の主体を区別した．ブログの書き手が主体の認知活動は『一人称心的操作』，それ以外の特定の主体の認知活動は『他人称心的操作』に分類される．さらに，「人と気軽にコミュニケーションが取れないのは大変だ」のように，主体は書き手とも取れるが，それ以外の主体とも取れる，明確に主体が定まらない記述は，『一般的心的操作』として区別した．また，佐藤（2007）[3]にならい，書き手やその他の主体の認知活動に当てはまらないものは，『外部知識』として区別した．

2.3.2 表現法分析

2.3.2.1 形態素解析

ブログ記事中で用いられる語彙の頻度分布を，広告の有無によって比較した．ChaSenを用いてブログ記事の形態素解析を行ない，統語カテゴリーごとに出現頻度を比較した．また，KH Coderを用いて，1）動詞，2）形容詞，3）形容動詞，4）副詞，5）接続詞の5つの統語カテゴリーについて，広告の有無による頻出語彙の比較を行なった．

2.3.2.2 表現方法分類

修辞的表現を比較するため，話題分類で使用した命題，総数2844命題に対して表現法タグを付与し，表現法タグ分類表を作成した（表3）．

表 3　表現法タグ分類表

表現法	表現法	
外向的表現	強調	単純強調
内向的表現		反語
省略		反復
伝聞		倒置
会話表現		体言止め
口語表現		感嘆表現

「～よ」「～ね」などの語尾を用いて，読み手に対する書き手の態度を表す表現を「外向的表現」とし，これに対して，ブログの読み手との対話を意識させない表現を「内向的表現」と定義した．また，ブログの書き手と他者との会話表現を「会話表現」，ため息や笑い声の表現などは「口語表現」に分類した．

3 結果

ジャンルを問わず，広告の有無によって異なる言語的特徴として，以下のような結果が得られた．

3.1 内容分析

3.1.1 話題分類

話題構成カテゴリーの各カテゴリーに分類された命題数を，広告の有無で比較すると，3ジャンルに一貫して，対象商品の話題が広告のある記事で有意に多く，対象商品から離れた話題が広告のない記事で有意に多くなった（$p<.01$，TVゲーム・対象から離れた話題のみ，$p<.05$）．広告のある記事ほど，話題が対象商品に集中しやすく，広告のない記事ほど，対象商品から離れた話題に変遷しやすいと言える．

3.1.2 認知要素分類

各認知タグが付与された命題の数を，広告の有無によって比較すると，3ジャンルに一貫して，「感性」タグに分類される命題の数に，有意差および有意な傾向が見られた．したがって，広告のある記事ほど，『グラフィックが凄すぎて発売当時から、いつかやるゲームとして決めていましたが、（広告有り／「ヴァルキ

表4　認知要素分類の統計的比較（*p<.05）

小分類	広告有り 命題数	広告無し 命題数
感性	書籍	
	29	11
	TVゲーム	
	45*	21*
	DVD	
	38*	32*

リープロファイル２」／google検索／B, 1.7）』のような，書き手の受けた商品の印象や雰囲気などの，感性的な表現が多くなると言える．

3.2 表現法分析

3.2.1 形態素解析

3.2.1.1 統語カテゴリー

各統語カテゴリーに出現する形態素数を比較すると，助動詞，接頭詞，名詞の3項目に，広告の有無による有意差が見られた．このうち，極端に出現数の少ない接頭詞は指標として扱いづらいと考えられるため，詳細な分析を避けた．また，各ジャンルはそれぞれ異なった話題構成を持っており，主題の影響を受けやすい名詞は，文章特徴の指標となりづらい．そこで，文章の主題と密接な関係がなく，文法的な機能や役割を果たす「機能語」として，文章特徴の指標となる助動詞について，さらに詳しい比較を行なった．

話題を「対象商品」に限定して，記事中で使用された基本形の出現数に対して，χ^2検定を行なったところ，「だ」「ます」という表現について，広告のある記事の方が有意に多くなった（$p<.01$）．

これらの助動詞は，『タイトルで明示されているように、「13歳のハローワーク」は数多の職業への「見出し」だ。（広告有り／「13歳のハローワーク」／google検索／A, 1.4-5）』のように，ブログの書き手の意見を文章の前面に押し出し，説得的に表現していると考えられる．すなわち，広告のある記事ほど，書き手の意見を強く表現した文が多いと見られる．

3.2.1.2 頻出語彙

5つの統語カテゴリーについて，広告の有無による頻出10語の比較を行なったところ，副詞を除く4つのカテゴリーにおいて，広告の有無による特徴が見られた．

3.2.1.2.1 動詞

動詞では，文章全体でも，話題を「対象商品」に限定した場合にも，広告のある記事で「感じる」が相対的に多くなった．

表5　動詞頻出語彙の出現数比較
（括弧内は抽出された動詞に対する出現割合）

動詞		広告有り	広告無し
抽出語	対象限定	出現数	出現数
感じる	―	13(2%)	9(1%)
	○	10(2%)	圏外

「感じる」という感性的な表現は，話題を限定すると，特に顕著に広告のある記事の中に頻出する．「感じる」は，『一点だけイマイチだなあと感じたのは、(広告有り／「メタルギアソリッド４」／Yahoo!検索／C, 1.55）』のような，主観的かつ感性的な評価の表現であり，書き手の受けた印象や，感じた雰囲気などを分類する，認知タグ「感性」の表すものと等価であると考えられる．

3.2.1.2.2 形容詞

形容詞では，広告のある記事ほど「良い」を多用し，広告のない記事ほど「面白い」を相対的に多く使用する傾向が見られた．この結果は，広告のある記事の書き手と広告のない記事の書き手の評価軸が，異なることを示している．広告のない記事で相対的に多用される「面白い」という感情は，『思ったより面白いな。（広告有り／「13歳のハローワーク」／Yahoo!検索／B, 1.2）』のように，個人的な体験に基づいて発生する評価である．これに対して「良い」という評価は，よ

表6　形容詞頻出語彙の出現数比較
（括弧内は抽出された動詞に対する出現割合）

形容詞		広告有り	広告無し
抽出語	対象限定	出現数	出現数
良い	―	21(13%)	11(8%)
	○	16(15%)	7(8%)
面白い	―	11(7%)	17(12%)
	○	11(10%)	12(13%)
凄い	―	8(5%)	圏外
	○	8(8%)	圏外

り一般的な表現であり，書き手は，自分が「面白い」と思わなくとも「良い」と表現することが出来る．例えば，『んで、肝心のゲームのほうですが、とりあえず『ソリッド』シリーズの最終作としては良い出来だと思います。(広告無し／「メタルギアソリッド４」／Yahoo!検索／C, 1.6-7)』という記述は，「良い」という評価を与えているが，書き手としては含みのあるような表現になっている．すなわち，広告のある記事ほど「良い」を多用して，書き手のもともと持っている評価とはバイアスのかかった評価，さらには書き手の評価とは異なる評価を，ブログ記事上でより多く表現している可能性が考えられる．

また，広告のある記事ほど「凄い」が多く使用されていた．「凄い」は強調の意味を含む，感性的な表現である．さらに，印象としての「凄い」は認知タグ「感性」に分類される語彙であり，広告のない記事で「凄い」が頻出語に入らないことは，すなわち広告のある記事ほど，感性表現，強調表現が多くなる傾向を支持していると言える．

3.2.1.2.3 形容動詞

形容動詞では，広告のある記事のみに「非常に」と変化させ，強調を表す語が現れた．また，話題を限定しない場合，ほぼ同数であった「好き」が，話題を限定すると，広告のない記事の方に相対的に多くなった（表7）．

表7　形容動詞頻出語彙の出現数比較
（括弧内は抽出された動詞に対する出現割合）

形容動詞		広告有り	広告無し
抽出語	対象限定	出現数	出現数
好き	—	19(10%)	20(11%)
	○	5(5%)	9(9%)
非常	—	5(3%)	圏外
	○	5(5%)	圏外

したがって，広告のある記事ほど強調表現を多用し，広告をつけない記事の書き手ほど「好き」という評価を用いる傾向が考えられる．

3.2.1.2.4 接続詞

接続詞では，話題の限定を問わず，広告のある記事ほど「でも」「しかし」を多用し，広告のない記事ほど「そして」を多用していた．そこで，接続詞頻出10語のうち，逆接にあたる接続詞（でも，しかし，いや，が，けれど，だが），および順接（そして，それから，しかも，だから）を，それぞれまとめてχ^2検定を行なった結果，広告のある記事は広告のない記事より，有意に逆接が多く，広告のない記事は広告のある記事より，有意に順接が多くなった（表8）．

表8　接続詞の統計的比較（*p<.05）

	広告有り	広告無し
抽出語	出現数	出現数
逆接	34*	24*
順接	12*	23*

3.2.2 表現方法分類

表現法タグの付与された命題数を，広告の有無で比較すると，強調の表現に有意な差が見られた（表9）．

表9　強調表現の出現数比較（**p<.01）

表現法		広告有り	広告無し
		命題数	命題数
強調	単純強調	4	2
	反語	5	0
	反復	1	3
	倒置	13	6
	体言止め	29	20
	感嘆表現	5	3
		57**	34**

細かい表現方法を比較すると，広告のある記事は反語や倒置を相対的に多く使用している．これらは意味の強調と同時に，読み手に与える印象も強調することになる．また，体言止めもその名詞を強調し，読み手に強い印象を与える．すな

わち，広告のある記事ほど，ブログの書き手が，読み手に与える印象を強調しようとする傾向を見ることができる．

4 考察

本研究によって得られた結果をまとめると，次のようになった．

a. 広告のある記事ほど，話題が対象商品に集中しやすく，広告のない記事ほど，対象商品から離れた話題に変遷しやすい．
b. 広告のある記事ほど，書き手の受けた印象など，感性的な表現が多くなる．
c. 広告のある記事ほど，書き手の意見を強調した表現をとる．
d. 広告のある記事ほど，強調表現を用いた印象の強調表現をとる．
e. 広告のある記事ほど，書き手は評価軸を複数持つ．

これらの結果を踏まえて，さらに深い考察を行う．

4.1 「評価の逆転」表現

広告のある記事で多く見られる逆接接続詞の前後関係を検討するため，「対象商品」話題の中で逆接接続詞を含む文，および逆接の助詞を含む文の，逆接を挟んだ前後の意味内容を確認した．すると，広告のある記事で14件，広告のない記事で5件，対象商品に対する「評価の逆転」表現が見られた（表10）．評価逆転の方向は，『好意的→非好意的』『非好意的→好意的』のどちらも見られているが，共通するのは，視点を変えることによる評価の変化である．表10の例では，逆接接続詞の前では，読み物として「面白い」と評価しているのに対し，接続詞のあとでは，参考書としての扱いとしては「どうかと思う」という，あまり好意的でない評価を取っている．このような「評価の逆転」は，広告のある記事の中で多く見られており，広告のある記事ほど，書き手が複数の視点で商品を評価することが伺える．

4.2 広告としてのブログ記事

雑誌とテレビの広告文の比較を行なった具（2004）[4]は広告文を構成する5つの情報を明らかにした（表11）．これらの情報は，本研究で行なった分析手法に全て対応しており，本研究の分析対象となったブログ記事は，広告の有無を問わず，具（2004）の示す情報を含んでいる（表12）．

表 11　具（2004）の広告文の構成情報分類

	具(2004)
1)	商品が必要になる背景的要素を表す状況情報
2)	商品の特性，品質などを表す品質情報
3)	購買者を説得するための表現である接近情報
4)	販売元の会社名など，直接的には商品に関係のない，周辺的な情報を表す関連情報
5)	商品名情報

表 12　具（2004）の分類に対応する本研究の分析指標

具(2004)		本研究
1)	⇒	認知要素「行為・状態」「状態・状況」
2)	⇒	認知要素「内容記述」「特徴言及」
3)	⇒	認知要素「推奨」，表現法「外向的表現」
4)	⇒	認知要素「情報記述」
5)		

したがって，商品を紹介・評価しているブログ記事は，「広告文」に近い構造を持つと言える．

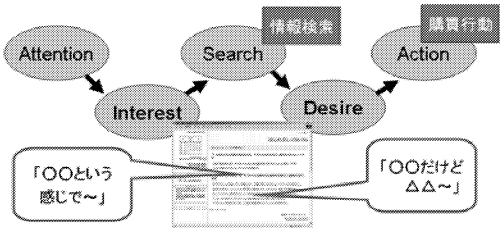
表 10　「評価の逆転」表現の例

見るには面白い。（好意的評価）　**でも、**
これを人生の参考にしなさいってすすめるのはどうかと思う。（非好意的評価）

（広告有り／「13歳のハローワーク」／Yahoo!検索／B，l.30-31）

これを踏まえて，ブログ記事を一種の「広告文」と考え，AISDAインターネット広告効果モデルと比較した．仁科（2007）[5]によれば，インターネットを介した広告の場合，消費者の広告接触から購買行動までの広告効果プロセスは，AISDAモデルで記述できるという．AISDAモデルとは，消費者が広告に接してから購買に至るまでに辿る，心理的・行動的プロセス「Attention→Interest→Search→Desire→Action」の5つのプロセスを表している．ブログ記事の場合「Attention」は，読み手がブログ記事を閲覧している時点で，ある程度達成している．そして，「Search」や「Action」はブログの読み手による行動であるので，ブログ記事が直接関与できる部分ではない．したがって，ブログ記事が効果をもたらすのは，商品情報の断片的な提供や，婉曲的な表現など嗜好を凝らして，商品に対するユーザーの興味を引く「Interest」，または情報収集することで“欲しい”と思わせる「Desire」の部分に当たると考えられる．したがって，ブログ記事が効果をもたらすのは，「Interest」または「Desire」の部分に当たると考えられる．ここで特に，広告を付与して記事を書いている書き手にとっては，読み手が「Interest」や「Desire」の次のプロセスに至ることは，自らの利益を生む可能性があるため，広告の付いたブログ記事が特に，読み手に期待する次の行動は「Search」，そして「Action」であると予測できる．したがって，広告のある記事とない記事では，「Interest」や「Desire」の喚起に，モチベーションの差が発生すると考えられる（表13）．このとき，感性的な表現の量や，強調表現を使うことによる読み手への印象の強調など，広告の有無によって見られた言語的特徴の差は，ブログの書き手の，読み手に対する「Interest」や「Desire」の喚起意欲の違いによって生じたものと考えられる．

表 13　インターネット広告効果モデル



5 参考文献

［1］Hancock, J.T.; Curry, L.E.; Goorha, S. & Woodworth, M.:「On Lying and Being Lied To: A Linguistic Analysis of Deception in Computer-Mediated Communication.」，*Discourse Process*，Vol.**45**，pp.1-23，2008.
［2］山下清美；川浦康至；川上善郎；三浦麻子：「ウェブログの心理学」，NTT出版，pp. 154-156，2005
［3］佐藤さやか；村井源；松本斉子；徃住彰文：「テキストの内容分析による文芸批評の認知構造の推定.」，第 10 回日本感性工学会大会発表論文集，2008.
［4］具軟和：「雑誌とテレビの広告文に関する一考察 -広告文の構成と語彙に注目して-」，御茶ノ水女子大学人間文化論叢，Vol. **7**，pp. 345-355，2004
［5］仁科貞文；田中洋；丸岡吉人：「広告心理学」，電通，pp. 51-111，2007

謝辞

本研究は科研費「知識共有のための価値指向型オントロジーの多分野多言語化」（20300074）の助成を受けた．

書評中の語を用いた図書に対する感性パラメータ自動設定の最適化

Optimization of automatic assignment of "Kansei" parameters to the books based on the words in online book reviews

原田 隆史[1*], 池内 正明[1,2]
Takashi HARADA, Masaaki IKEUCHI

*1：慶應義塾大学文学部図書館・情報学専攻
School of Library and Information Science, Keio University
〒108-8345 東京都港区三田 2-15-45
Email: ushi@slis.keio.ac.jp
2：現在の所属は，株式会社 安川電機

近年，人間の感性をもとにして小説を検索するシステムが開発されてきている。著者らもオンライン書評中の語を元として児童図書に感性パラメータの値を自動設定するシステムを開発してきた。このようなシステムでは，人間が感性パラメータとして積極的に高い値を付与した図書(特徴的な図書)をより選択的に検索できるような手法の開発がより望まれるが，従来の自動設定手法では特徴的な図書に対する値の付与は困難であった。そこで，本研究では，複数の書評中の語の選択手法および付与手法を組み合わせた実験を行い，どのような手法が特徴的な図書に対して最も適切にパラメータを付与できるかを比較した。図書 1425 冊分の書評を対象として感性パラメータの自動設定実験を行った結果，(1)再現率には書評中の語数が大きく影響している，(2)自動設定に用いる語数は 200 語から 300 語程度を用いたときに最も再現率が高くなる，(3)機械学習手法も重回帰分析をもとにした付与手法も再現率には大きな差はないことが明らかとなった。

In recent years Kansei retrieval systems have been developed to be practical use. We have developed the retrieval system using Kansei parameters based on online book reviews of children's books. There is a growing demand for Kansei retrieval systems that can retrieve the books whose value of Kansei parameters set high manually. Existing Kansei retrieval systems, however, cannot set appropriate value of Kansei parameters to those books featured by strong feelings, impressions or atmosphere, such as very amusing books or very sad. To resolve such problems, we experimented with various combinations of the methods to select the words in online book reviews and to set values of Kansei parameters in order to find the most appropriate combination to set parameters to those books. 1425 book reviews were used to assign the parameters automatically. The result showed that 1)the recall ratio was influenced considerably by the number of the words in book reviews, 2)the highest recall ratio was achieved when 200-300 words were used to assign parameters automatically, and 3)there was really not much difference between the approach of machine learning and multiple linear regression analysis.

キーワード：感性検索，感性パラメータ，書評，読書指導
Keyword: Kansei retrieval, Kansei Parameter, Online book revies

1. ニーズに対応した読書指導の必要性

近年，子どもに対する読書指導を行うなどの目的で，図書から感じる感情や雰囲気など幅広い図書の評価から検索することができるシステムの開発が進められてきている。その代表的なものとしては，英国の whichbook.net がある[1]。

著者らは，この whichbook をモデルにして，図書の雰囲気・図書から受ける印象を表す感性語をもとにした図書検索が可能な日本独自のシステムの試作を行ってきている[2)3)4)]。試作したシステムでは，人間の感情に関して「楽しい-悲しい」のような 12 種類の感性対を設定し(以下，「感性パラメータ」とする)，これに 1 から 5 までの値を付与し，これを用いて適切な図書を検索できるようになっている。感性パラメータに対する値の付与に際しては，人手で行う機能を用意するとともに，オンライン書評中の語を元に自動的に値を付与できる仕組みもそなえている。

この自動的に値を付与する手法としては，機械学習法および重回帰分析を用いた手法による実験を行い，感性パラメータによって異なるが第 1 表に示すような 30～70%の正解結果を得ることに成功している[2)3)]。しかし，現在採用している手法では，感性パラメータの値として五件法の中間である 3 点が付与されている図書に対する感性パラメータの評価値は正しく付与できることが多いが，1 点または 5 点の図書については，図書の多くについては正しい値の付与が困難である[3)]。

3 点という中間的な値は，その図書が何らかの特徴がない場合に付与されるケースが多い。読書指導という観点から考えた場合，評価値か 3 点という値を正しく付与された図書を正しく判定することよりも，1 点または 5 点というような強い感情に対応する値が付与された図書を正しく判別できることの方が実用性の高さにつながると考えられる。

そこで，本研究では「①ドキドキする-冷めた」および「⑩楽しい-悲しい」という2つの感性パラメータを対象に，書評中のどの語を選択し，ど

第1表　感性パラメータに対する付与実験の結果

感性パラメータ	書評中の語の種類	実験対象図書数	正解数(再現率)［機械学習手法］	正解数(再現率)［重回帰分析］
①ドキドキする-冷めた	23	1050	463(44.1%)	380(36.2%)
②びっくりする-平凡な	15	825	407(49.3%)	329(39.9%)
③心温まる-孤独な気持ち	51	1231	595(48.3%)	310(25.2%)
④夢の中の世界-現実的	61	1385	460(33.2%)	441(31.8%)
⑤明るい-恐い	17	709	345(48.7%)	280(39.5%)
⑥和む-イライラする	24	902	442(49.0%)	328(36.4%)
⑦爽やか-どんより	38	1146	464(40.5%)	375(32.7%)
⑧じーんとする-しらけた	34	1202	611(50.8%)	411(34.2%)
⑨元気が出る-がっかり	20	930	442(47.5%)	329(35.4%)
⑩楽しい-悲しい	73	1338	255(30.4%)	408(48.6%)
⑪笑える-深刻な	41	1195	515(43.1%)	364(30.5%)
⑫あっさり-じっくり	42	1221	556(45.5%)	434(35.5%)
⑬展開のはやい-ゆっくり	13	518	361(69.7%)	227(43.8%)

「原田隆史．感性語をもとにした読書支援システム．日本教育工学会研究報告集(学校図書館と情報教育)．Vol.2008, No.4, 2008, p.51-58」より

の付与手法を用いることが適切であるかについて比較実験を行うこととする。なお,「①ドキドキする-冷めた」については機械学習手法を用いた場合に,また「⑩楽しい-悲しい」については重回帰分析を用いた手法の場合に再現率が高い感性パラメータである。

2. 感性パラメータに対する値の付与

2.1 実験対象とする図書と書評

本研究で実験対象とする図書は,アマゾンジャパンの「こども」の中の「絵本」「児童図書」「ノンフィクション」「童話」と,「新書・文庫」の中の「ヤングアダルト」「こども」に含まれる図書23,487 冊から選択した。これらの図書に付与された,1)長さが短い書評,2)映画化,ドラマ化またはゲーム化された図書の書評,3)各種の受賞作品やベストセラーの書評,4)学習目的の本の書評などを除いて,なお3個以上の書評が存在する1,425 冊を実験対象とした。

図書の書評中の語のうち感性パラメータの値を付与する元とするものとしては,感情を示す語を選択的に抽出して用いた。たとえば,「白い」や「歩く」「食事」などの語は,人の感情や本の雰囲気を表す語とは言えないとしてあらかじめ除去している。本研究では書評を形態素解析ソフトウェア「茶筌」[5]にかけ,形容詞,形容動詞,副詞,動詞,名詞を抽出し,さらに人手によって抽出した語の中から人の感情や本の雰囲気を表す語の選択を行った。

その結果,1435 冊の図書の書評中に感性を表すと判断された語は1120 語であった。さらに,語幹が共通であるなどの同義語・類義語をまとめ,さらに出現回数 20 回未満の語句を除いた745 語を選定して用いた。

2.2 感性パラメータに対する値の付与手法

書評中の語を元にして感性パラメータに値を設定する手法としては,書評中のどの語をもとにするかについて三種類,値の付与手法として二種類の手法を組み合わせて実験を行った。

2.2.1 書評中の語の選択手法

書評中の語の選択に関しては,以下の手法を用いて語の選択を行い,それぞれ有効性について比較を行った。

1) 書評中の頻出上位の語を選択して用いる手法

学習データに用いる語数が多すぎる場合,教師データと学習データの間に不要な規則性を見出し,再現率が低下する可能性がある。そこで,745 語のうち頻出上位の 100 語,200 語,250 語,300 語,350 語,400 語,500 語,および745 語(出現回数 20 回以上のすべての語)を用いて値の付与実験を行い比較した。

2) 感性パラメータへの影響力が高いと思われる語を選択して用いる手法

本研究で用いる書評中の745 語のうち,感性パラメータにおける値の決定に影響力をもつ語は一部であり,それ以外の語は付与に有効ではない可能性が考えられる。そのため,各感性パラメータの値を決定するに際して影響力を与えている可能性の高い語を選択し,それらのみを用いた付与実験を行った。

各感性パラメータの値に影響力をもつ語としては,人手でパラメータに付与した1または5の図書に出現する比率が高い語を選択した。

3) 語の出現回数の正規化を行って語を選択する方法

図書ごとの書評数にはバラつきがあり,図書ごとの書評中に用いられる語の出現回数にも大きな差がある。一方,重複が図書の感性を表現するのに適切に機能している場合,正規化によって再現率が低下する可能性もある。

このような語の出現回数による影響を除外することが有効であるのか,逆に除外しない方がいいのかを検討するために,正規化処理を行って付与実験を行うこととした。具体的には,図書ごとに出現回数が 1 回以上の語の出現回数をすべて 1 とみなす前処理を行った上で付与実験を行った。

その際，頻出上位の語句を用いた実験でミクロ再現率とマクロ再現率の平均値がもっとも高かった手法に正規化を行って自動設定実験を行うこととした。

2.2.2 書評中の語を元にした値の付与手法

値の付与手法としては機械学習手法と重回帰分析を用いた手法の二つを用いた。

機械学習システムは，ワイカト大学で開発されたWekaを使用し[6]，判定器としては，決定木を用いた。

重回帰分析を用いた手法とは，人間によって感性パラメータに値が付与された図書を対象として，以下の線形重回帰式の各係数を推定し，推定された係数を元に未知データの値を計算する手法を意味する。感性パラメータの値は，1から5の整数値であり，ただし，計算された連続値をもとに値を付与するためには，1から5の値に区切る値(以下，閾値)を設定する必要がある。

$$y = a_1 x_1 + a_2 x_2 \ldots\ldots a_i x_i$$

ここでyは感性パラメータの値，
xは，書評中の語の出現回数を示す

この閾値の設定としては，1と5のような感情を強く示すものを優先的に正解させることを目的に，1または5の値から先に推定していく手法をとった。具体的には，まず1と5の値で再現率と精度が高くなるところ(再現率重視のF値を算出して仕様)を閾値とし，次に2と4の値の再現率と精度が高くなるところを閾値とする。最後に，残りを3の値とするという手法を採用した。 なお，重回帰分析はフリーの統計解析ソフトウェアであるRを用いた。

実験結果の評価指標としては，ミクロ再現率とマクロ再現率の両方を算出した。ミクロ再現率は，すべての図書のうち自動設定に正解した割合である。一方，マクロ再現率の算出方法はパラメータの値1から5の再現率をそれぞれ計算し，その平均をとるものであり，図書数が少ない1や5の値の重みを重視した指標となる。

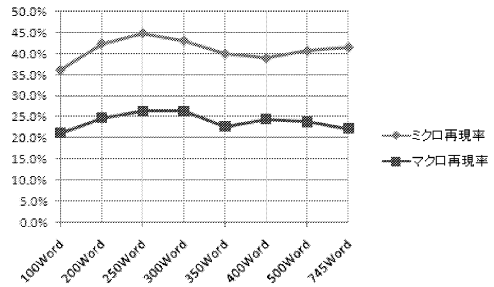
3. 語の選択および付与手法による比較

3.1 書評中の頻出上位の語を用いる手法

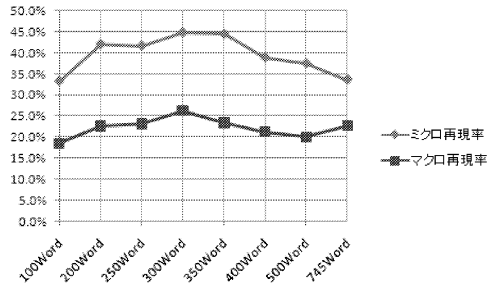
書評中に出現する語のうち，その出現頻度の上位から100,200,250,300,350,400,500語選択，および，異なる745語すべてを用いた場合の再現率を，第2表および第1図に示す。

第2表　書評中の頻度が上位である語を選択した場合の付与実験結果

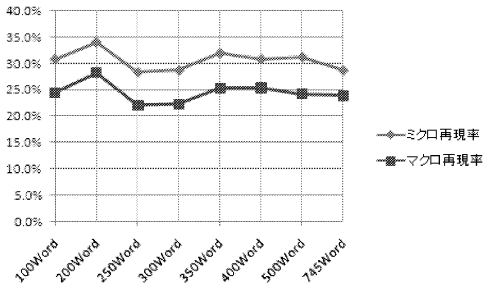
	①ドキドキする-冷めた				⑩楽しい-悲しい			
	ミクロ再現率		マクロ再現率		ミクロ再現率		マクロ再現率	
	重回帰分析	機械学習	重回帰分析	機械学習	重回帰分析	機械学習	重回帰分析	機械学習
100語	33.30%	36.10%	18.50%	21.20%	37.20%	30.90%	32.70%	24.50%
200語	42.10%	42.50%	22.60%	24.70%	31.60%	34.00%	27.80%	28.20%
250語	41.80%	44.90%	23.10%	26.20%	33.40%	28.40%	27.60%	22.10%
300語	44.90%	43.20%	26.20%	26.30%	33.70%	28.80%	26.20%	22.30%
350語	44.60%	40.00%	23.40%	22.70%	37.20%	31.90%	29.10%	25.30%
400語	38.90%	39.00%	21.20%	24.40%	37.90%	30.90%	32.00%	25.30%
500語	37.60%	40.70%	20.00%	23.80%	37.90%	31.20%	32.40%	24.20%
745語	33.70%	41.50%	22.70%	22.30%	28.80%	28.80%	25.90%	23.90%
1120語		42.80%		24.50%		35.90%		32.20%



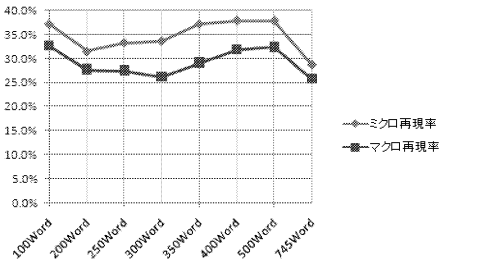
「①ドキドキする-冷めた」「機械学習」



「⑩楽しい-悲しい」「機械学習」



「①ドキドキする-冷めた」「重回帰分析」



「⑩楽しい-悲しい」「重回帰分析」

第 1 図　出現頻度語数ごとの再現率

第 2 表および第 1 図に示すように，いずれの感性パラメータおよび付与手法の場合にも，用いる語の数を変化させた場合に，再現率が1つもしくは2つの極大値を持つことが明らかとなった。ただし，重複する語も含めたすべての語を用いた場合と比較して，その差はあまり大きくなく，特に「⑩楽しい-悲しい」に対する値を付与する場合にはすべての語を用いた場合が最も再現率が高いという結果となった。

3.2　感性パラメータへの影響度が高いと思われる語を選択して用いる手法

感性パラメータへの影響度が高いと思われる語を選択した実験では，実験対象とした2つのパラメータについて，その値ごとの語の分布とは異なる語の分布を持つ語のみを対象として値の付与を行った。具体的には，「①ドキドキする-冷めた」の場合には「つまらない」「裏切る」「はらはら」「冒険」などの 23 語を，また，「⑩楽しい-悲しい」については「明るい」「殺す」「げらげら」「死」などの 73 語を異なる分布を持つ語として抽出して付与を行った。

付与実験の結果を第 3 表に示す。第 3 表に見られるように，特徴的と思われる語を選択して，それらの語のみを用いて感性パラメータの付与を行っても，マクロ再現率，ミクロ再現率ともにほとんどの場合で，語を選択せずに書評中に出現した 1120 語すべてを用いて付与を行った場合と比較して再現率が低い結果となった。

これは，特徴的な語が必ずしも感性パラメータの付与に有効な語であるとは限らず，付与の元となる語数が少ないことが逆に悪影響を与える可能性の方が高いことを示していると考えられる。

3.3　語の出現回数の正規化を行って語を選択する方法

各図書について同じ語が複数回出現した場合でも出現回数を1とカウントする正規化を行い，その上で機械学習および重回帰分析に基づく手法で値の付与を行う実験の結果を第 4 表に示す。

第 3 表　感性パラメータへの影響が強い語のみを用いて値の付与を行った結果

	①ドキドキする-冷めた			⑩楽しい-悲しい		
	語を選択する		しない	語を選択する		しない
	重回帰分析	機械学習	機械学習	重回帰分析	機械学習	機械学習
ミクロ再現率	41.40%	44.10%	42.80%	20.40%	32.40%	35.90%
マクロ再現率	21.50%	23.90%	24.50%	17.80%	25.50%	32.20%

第4表に見られるように，機械学習手法を用いた場合には正規化によって再現率が向上する傾向が見られた。ただし，正規化による再現率の工場は，それほど大幅なものではなかった。逆に，重回帰分析を用いる手法では，多くの場合でミクロ再現率およびマクロ再現率ともに低下した。これは，機械学習手法における正規化を行ったときの再現率の向上とは対照的な結果である。この原因としては，重回帰分析で用いた語の種類が，機械学習手法に比べて多いためか，あるいは重回帰分析を用いた手法と正規化を組み合わせることで悪影響が生じたことが考えられる。

4. まとめ

本研究は，書評中からの語の選択に関して三種類，書評中の語を元にした感性パラメータの付与に関して二種類の手法を組み合わせることで，より適切な感性パラメータの付与を行おうとしたものである。

図書 1425 冊分の書評を対象として感性パラメータの自動設定実験を行った結果，(1)再現率には書評中の語数が大きく影響していること，(2)自動設定に用いる語数は 200 語から 300 語程度を用いたときに最も再現率が高くなること，(3)機械学習手法も重回帰分析をもとにした付与手法も再現率には大きな差はないことが明らかとなった。今後，データ数を増やし語の選択や手法の検討を続けていきたい。

5. 引用文献

1) Whichbook.net. [2007-07-15] <http://www.whichbook.net/>

2) 桑田てるみ，原田隆史，小林加奈．感性語を検索語とした読書資料検索システムの開発に関する基礎研究-日本版 whichbook の作成．学校図書館学研究．日本学校図書館学会．Vol.8,2006,p.3-20

3) 原田隆史，江藤正巳，高柳知世．書評を用いた図書に対する感性パラメータの自動設定．情報知識学会誌．Vol.18，No.2，2008，p.153-160

4) 原田隆史．感性語をもとにした読書支援システム．日本教育工学会研究報告集(学校図書館と情報教育)．Vol.2008，No.4，2008，p.51-58

5) 形態素解析ソフトウェア「茶筌」．[2007-10-27] <http://chasen-legacy.sourceforge.jp/>

6) Weka 3 - Data Mining with Open Source Machine Learning Software in Java．[2007-08-06] <http://www.cs.waikato.ac.nz/ml/weka/>

第 4 表　正規化を行った上で値の付与を行った結果

	①ドキドキする-冷めた			⑩楽しい-悲しい		
	正規化を行う		行わない	正規化を行う		行わない
	重回帰分析	機械学習	機械学習	重回帰分析	機械学習	機械学習
ミクロ再現率	44.20%	44.90%	42.80%	25.30%	34.00%	35.90%
マクロ再現率	24.40%	26.30%	24.50%	29.00%	28.20%	32.20%

論文数・引用数からみたわが国の大学間格差の動向分析

A Trend Analysis on the Disparities among Japanese Universities Observed with the Number of Publications and Citation Counts

根岸正光*
Masamitsu NEGISHI*

*国立情報学研究所, 総合研究大学院大学
* National Institute of Informatics
〒101-8430 千代田区一ツ橋2-1-2
E-mail: negishi@nii.ac.jp

筆者は年来トムソン・ロイター社(通称 ISI)製の論文引用統計データベースに基づき, 大学・機関別の論文数・引用数関連の統計指標の研究開発を行っている. 2004 年の国立大学法人化以降, 各般の大学間競争助長政策のもと, 公私立大学を含めた大学間「格差」が拡がっているのではないかとの指摘が, 昨今なされるようになっている. 上記データベースにおける論文数統計からみた国立大学上位 8 大学とその他国立大との間での, 2006 年までの格差拡大の傾向については, すでに読売新聞記事「地方大, 揺らぐ研究基盤」の一部として公表したところである[1]. そこで本稿では, データを 2007 年までに更新した上で, 論文数のみならず引用数関連指標も加味し, またハーフィンダール係数(HHI)を算出して, 「格差」の年次的推移をさらに詳しく検証し, 今後の動向についての示唆を得るものとする.

The author has for years been developing statistical indexes on the number of papers and citations for universities and institutes based on the citation index databases of Thomson Reuters (ISI). Recently, there appear some discussions that the disparities among universities have been widened due to the competition encouraging policies by the government after the incorporatization of the national universities in 2004. The author has already publicized the trend of widened disparities between the top 8 national universities and the others up to 2006 in a newspaper article, "Local universities: Threatened research foundations" in The Yomiuri Shimbun. In the present paper, annual developments of the disparities are investigated in detail with the updated data of 2007 and indexes including the HHI: Herfindahl-Hirschman Index on citations as well as the number of papers, in order to get some idea on the future picture.

キーワード:研究評価, 引用, 成果主義, シミュレーション, HHI
Keywords: Research Evaluation, Citation, Performance-based System, Simulation, HHI

1　論文数・引用数による研究活動評価

大学・研究者の研究活動に対する「評価」は，専門研究者間での評価，すなわち同業者評価として従来から陰伏的に行われてきた．バブル以降の長引く不況のなかで，企業の研究開発投資余力が枯渇し，その代替物として大学を利用しようとの期待が高まり，いわゆる「改革」の一環として，2006 年の改正教育基本法に，大学の「社会貢献」が明記されるに至ったのは周知のとおりである．つまり大学に対する新たな評価軸として社会貢献度が加わったことになる．研究活動の社会的貢献度の評定は，文字どおり一般社会，すなわち非専門家の立場からの研究評価になるはずであるから実際上困難が大きい．しかしここに数量化された指標があれば，その大小比較は誰にでもできる．数量化指標が注目されるようになってきた背景にはこうした事情も作用しているのであろう．

研究活動に対する数量化指標として頻用されるのは論文数，(被) 引用数に関連した統計指標である．この場合，論文数とは専門家による査読審査に合格した一定水準の論文の数であり，引用数とはそうした論文中に，有用な先行研究として引用された回数であるから，その限りでいずれも専門家による同業者評価の集計結果である．結局，論文数・引用数関連指標も同業者評価に由来するものではあるが，得られた統計値に基づいて種々の立場，観点からの評価が可能になるのも事実であって，数量的指標の優位性の一端はここにあると考えられる．

表1 朝日新聞社「大学ランキング」誌への掲載状況

『大学ランキング』誌版	2002	2003	2004	2005	2006	2007	2008	2009	2010
発行年月	2001.5	2002.5	2003.5	2004.5	2005.4	2006.5	2007.5	2008.5	2009.5
標榜掲載大学数	682	699	710	717	726	723	733	737	?
調査対象期間	81-97	90-99	92-01	93-02	93-02	00-04	01-05	02-06	03-07
調査対象年数	17	10	10	10	10	5	5	5	5
調査対象論文数	740千	483千	549千	573千	573千	310千	370千	370千	371千
配列順序	論文数	総合引用度指数	総合引用度指数	総合引用度指数	総合引用度指数	案分加重引用度偏差値	加重引用度偏差値	加重引用度偏差値	加重引用度偏差値
引用度計算法	論文当引用数	同上	同上	同上	同上	同上	同上	同上	同上
掲載頁数	14	8	8	8	4	8	8	10 (*)	10 (*)
総合表掲載大学数	140	140	139	140	69	70	71	70	70
掲載分野数	24	20	20	20	10	20	20	20	20
NCR統計年次	81-97.6	81-99	82-01	81-02	81-02	81-04	96-05	97-06	98-07
NCR収録総論文数	85万	106万	120万	133万	133万	145万	83万	84万	86万

(*) 2009年版以降「引用力による交付金・補助金配分試算」、「引用当り単価ランキング」を追加

筆者は，従前からトムソン・ロイター社科学事業部（通称ISI）発行のNational Citation Reportを用いて，大学・機関別等の論文・引用関連指標の研究開発を行っており，その一端は朝日新聞社刊「大学ランキング」各年版で一般にも公表してきた<**表 1**>[2]．また，2005 年からは内閣府・総合科学技術会議での国立大学と産業総合研究所等，研究開発型独立行政法人 32 法人の年次評価の一環として，このデータベースによる統計調査研究を受託している[3]．

「大学ランキング」誌では，引用度（論文当たり引用数）の算定方式に改良を加

え，2008 年版以降「加重引用度偏差値」により配列している[4, 5]．最新 2010 年版での総合ランキング上位は<表 2>のようである[6]．

表2 加重引用度偏差値による総合ランキング(2003-2007年)

(Thomson Reuters (ISI), NCR for Japan, 1998-2007に対する根岸の調査結果。以下同様)

順位	大学・機関	論文数	加重引用度偏差値	順位	大学・機関	論文数	加重引用度偏差値
1	国立遺伝学研究所	598	169.5	16	首都大学東京	3,059	117.4
2	基礎生物学研究所	614	151.9	17	札幌医科大学	1,629	116.2
3	奈良先端科学技術大	1,796	127.3	18	星薬科大学	667	116.2
4	分子科学研究所	1,377	125.5	19	千葉大学	6,000	116.1
5	名古屋市立大学	2,065	125.0	20	滋賀医科大学	1,068	115.6
6	総合研究大学院大学	1,705	122.9	21	名古屋大学	13,178	115.1
7	京都薬科大学	765	122.5	22	東京女子医科大学	2,099	115.0
8	東京大学	34,042	121.8	23	東京工業大学	11,902	114.9
9	高エネ加速器研究機構	2,892	120.2	24	久留米大学	1,421	114.8
10	京都大学	24,881	120.0	25	杏林大学	790	114.8
11	大阪大学	21,126	119.9	26	川崎医科大学	777	114.7
12	大阪市立大学	4,042	118.9	27	奈良女子大学	911	114.6
13	生理学研究所	603	118.6	28	兵庫県立大学	1,498	114.4
14	東京医科歯科大学	3,683	118.3	29	熊本大学	3,581	114.3
15	横浜市立大学	2,253	117.6	30	筑波大学	8,562	114.0

2 案分論文数･引用数による予算配分･引用単価

納税者評価—前記の大学の社会貢献と関連させ，また成果主義的予算配分の潮流に鑑み，昨年発行の「大学ランキング 2009 年版」には，国私立大への交付金・助成金の引用力に応じた配分試算，シミュレーションを行い，また引用 1 件当りの単価計算を試みた[7]．これらは，いわば納税者的観点での評価ともいうべきもので，研究活動とそれを支える税金額との関連にひとつの示唆を与えるものであろう.本年発行の同誌 2010 年版では，データを更新するとともに計算法を改変した[6]．紙数の関係で割愛した部分や「下位」大学の状況を含め，改めて全体を詳細にみると次のようである.

引用力に応じた予算配分—引用力の指標としては，ごく最近の状況を反映させるために，2003～2007 年の 5 年間に刊行された論文に対する 2006～2007 年の 2 年間の引用回数に限って調査統計した．また，共著論文については，共著機関（大学・企業・外国機関等）数に応じてこれを案分計算した上で，大学別に集計している．この「案分引用数」の各国私大別のシェアに応じて予算配分を計算し，実績値との増減をみたのが<表 3>である.

国立大には運営費交付金と称する経費予算が文部科学省から毎年措置されている（2007 年度 1 兆 1,115 億円）．私立大には経常費補助金が同じく文部科学省から配分されている（同 3,003 億円）．私大方面からは，私大への補助金をさらに手厚くし，国立大とのイコール・フッティングを確保するべしとの意見もあるようなので，ここでは運営費交付金と経常費補助金の合計額 1 兆 4,118 億円を一括して上記案分引用数に応じて配分計算した.

表3　引用力による交付金・補助金の配分試算（増減額順）

（2003-2007年の論文に対する2006-2007年における引用数を共著大学別に案分）
（2007年度国立大運営費交付金と私立大経常費補助金の合計を上記案分引用数に応じて再配分）

	大学名	交付金・助成金実績（百万円）	引用力準拠増減額（百万円）	増減%
	国立大学計	1,111,501	+6,551	+0.6
	私立大学計	300,272	−6,551	−2.2
	国私大計	1,411,773	0	0.0
1	東京大学	89,943	+62,372	+69.3
2	京都大学	60,874	+61,851	+101.6
3	大阪大学	51,846	+48,117	+92.8
4	東北大学	51,899	+25,970	+50.0
5	東京工業大学	22,232	+21,365	+96.1
6	名古屋大学	34,655	+16,939	+48.9
7	慶應義塾大学	8,734	+16,189	+185.4
8	北海道大学	41,192	+10,662	+25.9
9	東京理科大学	3,244	+9,915	+305.6
10	九州大学	46,963	+8,908	+19.0
11	順天堂大学	3,969	+5,471	+137.9
12	千葉大学	19,045	+5,429	+28.5
13	東京女子医科大	3,142	+4,782	+152.2
14	総合研究大学院	1,973	+4,377	+221.8
15	自治医科大学	2,635	+4,212	+159.8
16	京都薬科大学	513	+3,930	+765.8
17	岡山大学	19,476	+3,836	+19.7
18	昭和大学	4,072	+3,784	+92.9
19	北里大学	4,108	+3,617	+88.0
20	奈良先端科技大	6,424	+3,525	+54.9
21	帝京大学	1,421	+3,355	+236.2
22	東京薬科大学	922	+3,351	+363.5
23	兵庫医科大学	1,503	+3,181	+211.6
24	日本医科大学	2,645	+3,146	+118.9
25	久留米大学	2,662	+3,034	+114.0
26	東邦大学	2,218	+2,945	+132.7
27	星薬科大学	409	+2,920	+713.6
28	早稲田大学	9,312	+2,438	+26.2
29	関西医科大学	1,749	+2,384	+136.3
30	近畿大学	5,258	+2,375	+45.2
31	獨協医科大学	1,942	+2,342	+120.6
32	東京農工大学	6,619	+2,115	+32.0
33	東京医科大学	2,084	+1,891	+90.8
34	東海大学	5,944	+1,775	+29.9
35	川崎医科大学	904	+1,559	+172.5
36	名古屋工業大学	5,239	+1,556	+29.7
37	福岡大学	3,686	+1,492	+40.5
38	埼玉医科大学	2,388	+1,477	+61.9
39	藤田保健衛生大	1,821	+1,428	+78.5
40	杏林大学	1,632	+1,346	+82.5
41	金沢医科大学	1,120	+1,312	+117.1
42	神戸薬科大学	278	+1,289	+463.4
43	豊田工業大学	489	+1,272	+260.2
44	大阪医科大学	1,302	+1,259	+96.7
45	明治薬科大学	466	+1,191	+255.7
46	神奈川大学	1,781	+1,153	+64.7
47	東北薬科大学	412	+1,103	+268.0
48	日本歯科大学	0	+1,063	――
1	筑波大学	43,395	−16,558	−38.2
2	鹿児島大学	16,524	−8,473	−51.3
3	琉球大学	13,419	−8,354	−62.3
4	東京学芸大学	9,113	−8,133	−89.2
5	北海道教育大	7,549	−7,209	−95.5
6	弘前大学	11,780	−6,552	−55.6
7	鳥取大学	12,066	−5,815	−48.2
8	大阪教育大学	6,534	−5,785	−88.5
9	一橋大学	5,862	−5,613	−95.7
10	宮崎大学	10,500	−5,440	−51.8
11	山形大学	12,281	−5,371	−43.7
12	佐賀大学	11,173	−5,318	−47.6
13	島根大学	10,807	−5,282	−48.9
14	新潟大学	18,470	−5,278	−28.6
15	香川大学	10,684	−5,194	−48.6
16	神戸大学	22,955	−4,948	−21.6
17	愛知教育大学	5,214	−4,854	−93.1
18	愛媛大学	13,897	−4,801	−34.5
19	秋田大学	9,969	−4,790	−48.0
20	信州大学	16,152	−4,752	−29.4
21	岐阜大学	14,224	−4,749	−33.4
22	東京芸術大学	4,670	−4,642	−99.4
23	富山大学	14,132	−4,640	−32.8
24	岩手大学	7,608	−4,557	−59.9
25	高知大学	10,285	−4,505	−43.8
26	大分大学	9,461	−4,344	−45.9
27	茨城大学	7,449	−4,341	−58.3
28	長崎大学	17,322	−4,278	−24.7
29	宇都宮大学	6,073	−4,205	−69.2
30	福井大学	9,752	−4,174	−42.8

イコール・フッティング―ここまず興味深いのは，こうしたイコール・フッティング的配分によっても，国立・私立間の配分にはほとんど変動がなく，むしろ国立の方に 66 億円程を振り替えるという計算になったことである（<**表 3**>大学計欄参照）．すなわち，国立・私立間では図らずも既に引用力に応じた資金配分になっており，従って実績値からの配分額の移動は，国立大，私立大それぞれの内部での一部大学への集中化ということになる．なお，国立・私立間での資金配分実績値が引用力に現実に比例しているということは，金をかければそれ相応の研究成果が得られるとの解釈も成り立つわけで，従って，国立から私立への資金振替えを積み増せば，その分私大側の引用が増えるという可能性も考えられよう．

増額大学―<**表 3**>では実績値からの引用力比例による増減額の順に大学を配列してみた．最大は東大の 624 億円贈で，京大もほぼ同額である．ただしこれを増加率でみると，東大+69.3%，京大+101.6%（倍増）となり，京大の方が引用力が強

いことが分かる．増加率では，大規模私大の+100%超が目立つが,その多くが医薬系単科大学である．

減額大学—表の後半は「削減額」順のリストであり，筑波大の 166 億円減（-38.2%）が抜群である．これは大規模大学への集中化傾向の中では異質であり，研究分野の構成に立ち入ってさらにその要因を検討する必要があろう．また ISI 収録対象の英文論文が少ない教育系・文化系大学では 90%の削減という，非現実的ともいいえる数字になるのも注目される．これはむしろ，成果主義的予算の効果と副作用の双方を検証できるという点で，こうした試算の有用性，必要性を示すものと考えるべきではなかろうか．

表4　引用の単価：案分被引用数１件あたりの交付金・補助金＋科研費取得額

（2003-2007年の論文に対する2006-2007年における引用数を共著大学別に案分）
（2007年度国立大運営費交付金、私立大経常費補助金と取得科学研究費の合計を上記案分引用数で除した）
（案分引用数500件以上の国私大について「単価」順に掲載）

順位	大学	引用当り交付金補助金科研費（千円）	2007年度交付金・補助金＋科研費（百万円）	案分被引用数	順位	大学	引用当り交付金補助金科研費（千円）	2007年度交付金・補助金＋科研費（百万円）	案分被引用数
1	京都薬科大学	343	617	1801.2	26	昭和大学	1,377	4,385	3185.4
2	星薬科大学	345	466	1349.8	27	北里大学	1,460	4,571	3131.8
3	神戸薬科大学	507	322	635.4	28	愛知医科大学	1,469	1,237	841.6
4	東京薬科大学	649	1,124	1732.6	29	杏林大学	1,479	1,785	1207.5
5	東京理科大学	730	3,893	5335.4	30	京都大学	1,510	75,155	49758.8
6	東北薬科大学	734	451	614.3	31	藤田保健衛生大	1,518	2,000	1317.3
7	明治薬科大学	736	495	671.9	32	大阪大学	1,541	62,476	40529.7
8	豊田工業大学	773	552	713.7	33	東京工業大学	1,544	27,296	17676.2
9	総合研究大学院	800	2,060	2574.6	34	岡山理科大学	1,599	1,165	728.5
10	帝京大学	840	1,626	1936.5	35	徳島文理大学	1,601	1,344	839.5
11	兵庫医科大学	940	1,784	1899.1	36	埼玉医科大学	1,615	2,531	1566.9
12	北陸大学	985	518	525.9	37	神奈川大学	1,621	1,928	1189.5
13	川崎医科大学	1,006	1,005	998.7	38	名城大学	1,736	1,729	995.9
14	自治医科大学	1,059	2,940	2776.2	39	東京大学	1,800	111,165	61755.9
15	東京女子医科大	1,073	3,447	3212.7	40	学習院大学	1,826	1,414	774.1
16	慶應義塾大学	1,101	11,121	10104.9	41	奈良先端科技大	1,839	7,419	4033.8
17	東邦大学	1,133	2,372	2093.2	42	東京歯科大学	1,843	1,205	653.7
18	順天堂大学	1,138	4,357	3827.5	43	近畿大学	1,862	5,761	3094.4
19	関西医科大学	1,139	1,909	1675.7	44	福岡大学	1,873	3,933	2099.5
20	獨協医科大学	1,174	2,039	1737.1	45	北海道医療大学	1,904	1,172	615.8
21	日本医科大学	1,246	2,927	2348.1	46	東北大学	1,974	62,310	31571.8
22	金沢医科大学	1,257	1,239	986.1	47	名古屋大学	1,974	41,288	20918.8
23	久留米大学	1,302	3,007	2309.1	48	東京慈恵会医大	2,012	3,520	1749.1
24	東京医科大学	1,362	2,196	1611.9	49	聖ﾏﾘｱﾝﾅ医科大	2,058	2,370	1151.7
25	大阪医科大学	1,369	1,422	1038.5	50	東海大学	2,078	6,502	3129.5

引用の「単価」—次に，国費・税金による大学の資金として，運営費交付金，経常費補助金に文部科学省科学研究費補助金取得額を加えたものを，上述の案分引用数で割って，1 引用当りの「単価」を算出してみた<**表 4**>．前回統計[5]に比べると，引用の計測期間を 5 年から 2 年に短縮したため，1.8 倍見当単価数値が上昇していおり，第 1 位の京都薬科大が 1 引用当り 34 万円,東大は 180 万円といった数字になる．

私立の薬科大学が上位にくることで，引用獲得という観点では，薬学における研究費の資金効率が非常に高いということがわかり，これに続くのが私立の医科大学となる．こうした中にあって，東京理科大，豊田工大が「健闘」しているのが注目される．国立大学は文字通り，運営費交付金を含めた国費負担率が高く，単価も高くなる．その中で，京大，阪大，

東工大が 150 万円前後の数字で上位に来るが，私立総合大学の慶応大の 110 万円とはさすがに開きがある．なお，国立大学の中では総合研究大学院大学が上位に来ているが，この大学は大学共同利用機関の研究所を基盤機関とする特殊な構造の大学であって，単純な比較はできない．

すでに別途シミュレーションにより検討したとおり，引用力に準拠した予算配分を継続反復適用していった場合，資金効率の高い京都薬科大学に次第に予算を集中していく結果となる[5]．こうした点にも鑑み，ともあれこの際，前記の加重引用度偏差値のように，大学の研究分野構成を加味した単価指標を開発することにより，研究水準と資金効率の関係をより明確に表現できるよう検討してみたい．

3 国立大学間「格差」の動向

国立大学の法人化以降，国立大間の格差が各側面で拡大しているのではないかとの指摘がなされるようになっている．

表5 国立大学における論文数・引用数の集中の状況

（1998-2007年の合計値）

	大学	論文数	案分論文数	引用数	案分引用数		大学	論文数	案分論文数	引用数	案分引用数
1	東京大	65,081	34,800.1	837,242	389,798.2	44	奈良先端大	3,156	1,736.1	49,709	22,415.1
2	京都大	46,916	27,569.9	579,246	314,005.2	45	山梨大	3,123	1,943.9	28,652	15,214.7
3	大阪大	40,546	23,716.2	516,384	260,555.4	46	九州工大	3,015	1,838.6	14,312	7,877.6
4	東北大	38,704	22,695.1	380,691	196,273.2	47	香川大	2,893	1,787.3	21,185	12,211.5
5	九州大	26,526	16,421.4	246,630	140,596.7	48	豊橋技科大	2,776	1,767.4	16,866	10,720.9
6	北海道大	25,742	15,716.4	225,616	126,205.8	49	埼玉大	2,738	1,555.7	18,920	8,793.0
7	名古屋大	25,568	14,304.7	269,817	133,205.7	50	浜松医科大	2,727	1,580.2	28,280	14,361.8
8	東京工大	22,727	14,106.0	208,147	111,086.5	51	茨城大	2,704	1,509.1	14,481	6,944.2
9	筑波大	16,642	8,943.4	161,569	68,668.3	52	総合研究大	2,682	977.2	35,175	11,484.1
10	広島大	14,987	8,841.4	127,599	66,377.9	53	大分大	2,635	1,849.7	19,812	12,815.1
11	岡山大	12,264	7,502.5	103,582	56,034.6	54	岩手大	2,404	1,360.2	12,786	6,433.5
12	千葉大	11,535	6,635.5	119,102	58,121.1	55	滋賀医科大	2,287	1,323.0	24,642	13,134.5
13	神戸大	10,518	6,055.9	102,262	48,286.7	56	北陸先端大	2,171	1,384.6	16,296	10,220.0
14	金沢大	8,413	4,937.9	87,605	43,806.2	57	長岡技科大	2,166	1,486.4	12,589	8,051.5
15	新潟大	8,323	4,742.0	88,139	37,051.3	58	東京海洋大	1,729	1,025.7	10,615	5,424.9
16	熊本大	7,213	4,089.1	86,162	40,861.6	59	旭川医科大	1,671	952.1	19,642	9,575.5
17	東京医歯大	7,086	3,793.7	93,897	42,210.1	60	奈良女子大	1,611	745.8	13,263	4,187.9
18	信州大	6,980	4,181.7	61,750	30,858.8	61	お茶の水大	1,442	789.2	10,636	4,946.0
19	長崎大	6,792	4,125.3	62,020	33,252.9	62	宇都宮大	1,427	840.2	7,291	3,710.4
20	岐阜大	6,561	3,851.8	57,912	27,040.6	63	帯広畜産大	1,305	682.5	7,866	4,011.3
21	群馬大	6,548	4,093.0	58,332	31,379.0	64	室蘭工大	1,031	628.4	4,473	2,487.6
22	徳島大	6,545	3,866.0	60,271	32,024.0	65	北見工大	763	505.1	3,835	2,487.5
23	富山大	5,810	3,559.7	45,330	25,337.9	66	東京学芸大	701	359.8	3,732	1,464.5
24	山口大	5,801	3,726.2	42,612	25,162.8	67	大阪教育大	672	402.5	4,286	2,042.5
25	鹿児島大	5,586	3,308.2	46,035	23,662.8	68	和歌山大	603	332.6	3,681	1,810.8
26	東京農工大	5,271	3,157.1	39,586	21,284.9	69	愛知教育大	445	198.4	4,201	1,604.8
27	愛媛大	5,008	3,013.9	45,761	24,201.5	70	一橋大	387	252.3	1,291	736.9
28	三重大	4,945	3,073.3	45,307	24,839.5	71	北海道教大	293	156.0	1,689	904.9
29	山形大	4,819	2,833.2	35,770	18,012.3	72	奈良教育大	289	162.2	1,849	932.1
30	静岡大	4,490	2,787.3	30,312	16,510.8	73	鳴門教育大	267	135.2	1,244	550.4
31	名古屋工大	4,158	2,799.4	21,292	14,386.3	74	福岡教育大	248	117.2	3,406	1,190.7
32	鳥取大	4,086	2,484.2	34,889	18,734.6	75	上越教育大	237	103.3	3,677	1,170.7
33	佐賀大	3,848	2,297.9	32,098	15,524.7	76	宮城教育大	211	63.7	3,807	594.6
34	福井大	3,680	2,211.6	29,713	15,239.3	77	兵庫教育大	198	103.6	1,048	455.9
35	横浜国立大	3,614	2,247.6	22,233	13,064.7	78	福島大	177	81.2	835	249.5
36	高知大	3,596	2,191.2	26,990	15,594.6	79	京都教育大	172	57.3	1,095	264.9
37	島根大	3,574	2,268.8	24,974	14,912.4	80	筑波技術大	164	74.5	887	354.2
38	宮崎大	3,447	1,915.3	34,906	16,233.6	81	鹿屋体育大	124	48.4	858	278.8
39	京都工繊大	3,400	2,038.2	22,664	12,434.4	82	滋賀大	104	53.1	576	225.2
40	琉球大	3,306	1,899.6	23,867	11,347.0	83	小樽商科大	78	57.9	242	169.3
41	秋田大	3,228	2,073.2	22,336	12,029.3	84	政策研究大	67	42.3	235	139.2
42	弘前大	3,204	1,874.8	26,178	13,011.4	85	東京芸術大	40	27.2	124	65.4
43	電気通信大	3,197	1,911.0	18,700	8,896.9	86	東京外語大	22	12.9	60	26.2

研究成果の重要な指標である論文数，引用数ではどうなっているか，以下この点を分析する．まず国立 86 大学における 1998～2007 年の 10 年間における論文数，引用数を案分数値を含めて統計すると<**表 5**>のようになる．

上位 8 大学との間の「落差」— 一瞥して旧 7 帝大に東工大を加えた 8 大学とそれ以下の間には，既に大きな落差が生じていることがわかる．すなわちこの 8 大学で論文数，案分論文数，引用数，案分引用数において，それぞれ全体の 51.7%，51.3%，57.9%，58.4%という大きな部分を占めるのである．ここで，論文数＞案分論文数，引用数＜案分引用数のように，単純（重複）集計値と案分値のシェアの大小が逆転しているのは，8 大学とその他大学間の共著論文が相対的に引用度が高いということを示唆していると考えられる．

案分計算での目減り—ちなみに，案分計算での「目減り」の割合を計算し，論文数や引用数で目減りの大きい大学と小さい大学を取り出すと<**表 6**>のようである．教育系大学は論文数も少なく，その上他大学との共著が多いことがうかがわれる．引用数の大きな目減りは，大学単独論文では多くの引用が得られないということであるから，これは大学の「自立性」の問題にも関連して来よう．一方目減りが少ない大学には文科系大学や工科大，技科大が現れるのも興味深い点で，独自性の強い「単独行動主義」であるというべきか．この辺り，具体的な共著相手の詳細に立ち入って調査するとさらに興味深い結果が得られると思われる．また，前出の「減額大学」の多くが，目減り率ではその大小方向に二極分化するのも注目されるところであろう．

表6　共著案分による目減り率

（表5における案分論文数/論文数、案分引用数/引用数の%表示）

目減りの大きい大学				目減りの小さい大学			
	大学	論文数	引用数		大学	論文数	引用数
1	宮城教育大	30.2	15.6	1	小樽商科大	74.3	70.0
2	京都教育大	33.3	24.2	2	大分大	70.2	64.7
3	総合研究大	36.4	32.6	3	長岡技科大	68.6	64.0
4	鹿屋体育大	39.0	32.5	4	東京芸術大	67.9	52.8
5	上越教育大	43.6	31.8	5	名古屋工大	67.3	67.6
6	愛知教育大	44.6	38.2	6	北見工大	66.2	64.9
7	筑波技術大	45.4	39.9	7	一橋大	65.2	57.1
8	福島大	45.9	29.9	8	山口大	64.2	59.1
10	奈良女子大	46.3	31.6	9	秋田大	64.2	53.9
11	福岡教育大	47.2	35.0	10	北陸先端大	63.8	62.7
12	滋賀大	51.1	39.1	11	豊橋技科大	63.7	63.6
	東京大	53.5	46.6	12	島根大	63.5	59.7
	京都大	58.8	54.2	13	政策研究大	63.2	59.2

上位 8 大学との格差の拡大傾向—前記の上位 8 大学とその他大学間の論文数，引用数とそのシェアの年次的推移を追うと<**表 7**>のようであり，シェアをグラフ化したものが<**図 1**>，<**図 2**>である．単純集計の論文数の 2006 年までの格差動向については，すでに読売新聞紙上で公表したところであるが，上記図表ではデータを 2007 までに更新し，引用数と案分値も計算して示した[1]．

表7 国立大学における論文数・引用数の集中化の動向

（上位8大学：旧7帝大+東工大とその他78国立大の間の格差の年次推移）

（2007、2006年における論文数の落ち込みはデータベースへの収録遅れに起因するもので通例）

		1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
論文数	上位8大学	25,696	27,413	28,166	28,090	29,209	30,198	31,788	31,678	31,742	27,830
	その他	24,716	25,963	26,814	27,137	27,605	28,614	28,747	28,901	28,740	25,193
	国立大計	50,412	53,376	54,980	55,227	56,814	58,812	60,535	60,579	60,482	53,023
案分論文数	上位8大学	16,201	17,054	17,179	16,728	17,024	17,300	17,873	17,572	17,483	14,917
	その他	15,767	16,324	16,693	16,517	16,421	16,786	16,383	16,210	15,759	13,581
	国立大計	31,968	33,379	33,872	33,245	33,444	34,086	34,256	33,782	33,242	28,498
論文数%	上位8大学	*51.0*	*51.4*	*51.2*	*50.9*	*51.4*	*51.3*	*52.5*	*52.3*	*52.5*	*52.5*
	その他	*49.0*	*48.6*	*48.8*	*49.1*	*48.6*	*48.7*	*47.5*	*47.7*	*47.5*	*47.5*
案分論文数%	上位8大学	*50.7*	*51.1*	*50.7*	*50.3*	*50.9*	*50.8*	*52.2*	*52.0*	*52.6*	*52.3*
	その他	*49.3*	*48.9*	*49.3*	*49.7*	*49.1*	*49.2*	*47.8*	*48.0*	*47.4*	*47.7*
引用数	上位8大学	483,718	479,556	484,051	431,695	400,854	339,467	289,216	203,606	115,745	35,865
	その他	352,546	364,558	327,404	330,093	284,985	252,486	206,555	146,189	81,254	25,834
	国立大計	836,264	844,114	811,455	761,788	685,839	591,953	495,771	349,795	196,999	61,699
案分引用数	上位8大学	264,831	262,735	254,654	220,942	198,559	165,396	139,387	96,141	53,210	15,873
	その他	189,263	193,401	173,757	163,100	140,101	120,593	95,825	65,959	36,386	10,751
	国立大計	454,095	456,136	428,410	384,042	338,659	285,989	235,212	162,100	89,595	26,623
引用数%	上位8大学	*57.8*	*56.8*	*59.7*	*56.7*	*58.4*	*57.3*	*58.3*	*58.2*	*58.8*	*58.1*
	その他	*42.2*	*43.2*	*40.3*	*43.3*	*41.6*	*42.7*	*41.7*	*41.8*	*41.2*	*41.9*
案分引用数%	上位8大学	*58.3*	*57.6*	*59.4*	*57.5*	*58.6*	*57.8*	*59.3*	*59.3*	*59.4*	*59.6*
	その他	*41.7*	*42.4*	*40.6*	*42.5*	*41.4*	*42.2*	*40.7*	*40.7*	*40.6*	*40.4*

ここではまず，<図 1>にみるとおり2004 年の国立大学法人化と同時に論文数における両者間の格差が広がり，それが以後年々大きくなっているようにみえる．法人化の「効果」が即座に現れた格好であって，まさに驚くべき結果ではなかろうか．また，単純集計値と案分値の差は 2004 年以降小さくなる傾向がみられるが，これは，大学単独論文，あるいは 8 大学内での共著が増えた結果と考えられ，単独行動主義や上位グループ内での集団行動が盛んになっているということを示唆している．

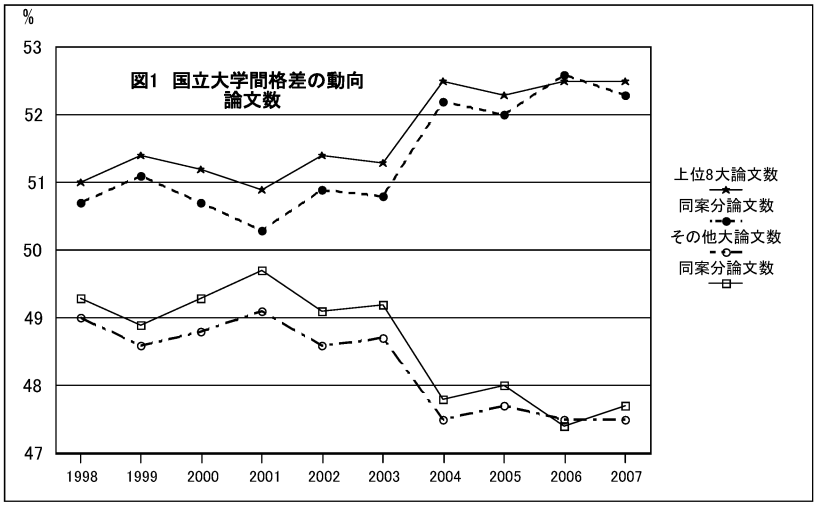
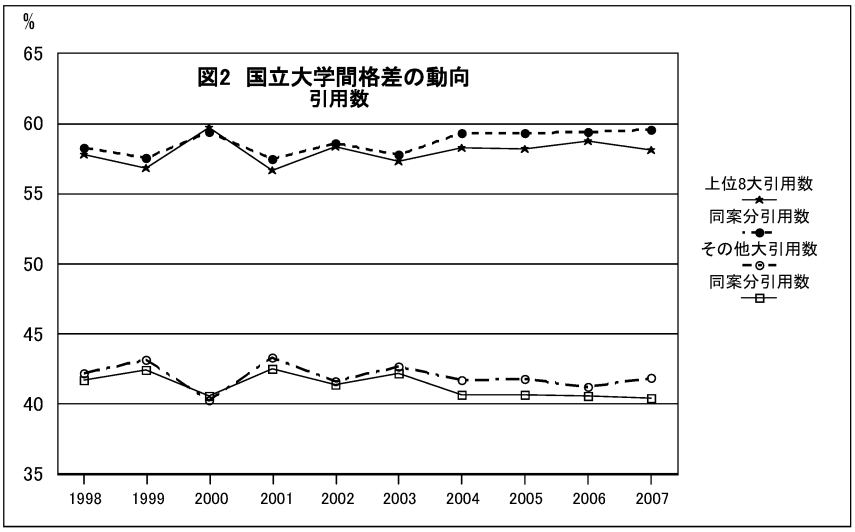


図1 国立大学間格差の動向
論文数

次に<図 2>に従い引用数の動向をみると，そもそも論文数に比して引用数は上位 8 大学への集中度が高く，6:4 に近い比率になっている．また，論文数と同様に 2004 年以降について，それなりの格差拡大が現れている．引用数については，

その蓄積，安定に一定の年数を要するので，にわかに判定できない部分もあるが，論文数と同様の傾向にあることは間違いなかろう．単純集計値と案分値の差をみると，2004 年以降にそれが拡大し，8 大学では，論文数とは逆に，案分引用数シェアの方が引用数シェアを相当上回るようになっているから，高引用論文の上位 8 大学への集中が進んでいるのであろう．

以上を要するに，2004 年の法人化以降，論文の量と質(引用数)の両面において，上位 8 大学の優位性がさらに増強されているとみてよいであろう．



4 HHI(ハーフィンダール・ハーシュマン指数)による集中度の動向分析

前項では上位 8 大学とその他大学という「勝ち組：負け組」式の二分法での対比を行ったが，全体を見渡しての集中度の趨勢はどうか．この点をみるために，独占禁止政策で多く用いられるハーフィンダール・ハーシュマン指数(Herfindahl-Hirschman Index, HHI)を用いて分析する．HHI は，各参加者の市場占有率の%表示の 2 乗値を累計したもので，1 社独占の場合は 100^2 で 10,000，2 社均衡なら 50^2*2=5,000，4 社均衡なら 25^2*4=2,500，占有率 40%，25%，20%，10%，5%の 5 社がいる場合は，$40^2+25^2+20^2+10^2+5^2$=2,750 が HHI であり，0<HHI<=10,000 の値をとる．

HHI は，2007 年の公正取引委員会の独占禁止法運用指針の改正において，市場の集中度，寡占度の判定基準としてさらに重視されるようになり，注目されている[8]．分布の集中度，不平等度を表す指標としてはジニ(Gini)係数が用いられることも多い．これは参加者数に関係なく全員平等であれば 0 となるが，HHI は上例のように各者同率でもその数によって，参加者が多い程低い値になる．また分布の形状(ローレンツ曲線)の異なりも HHI の方がよく反映される．

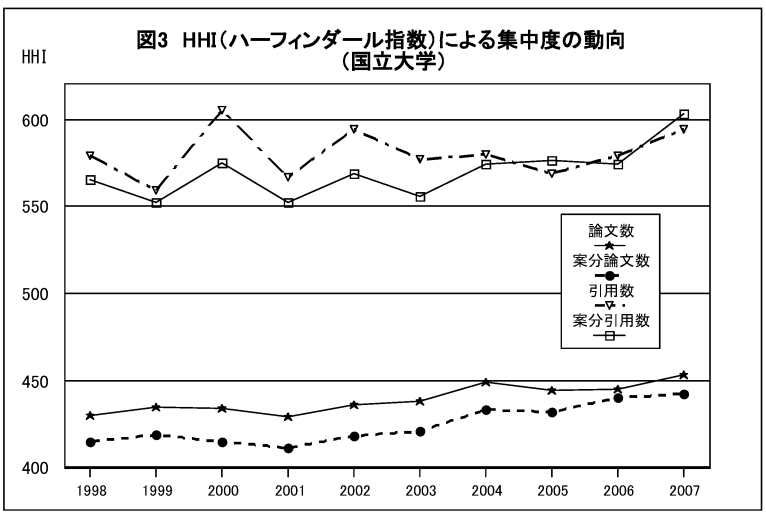
表8　HHI（ハーフィンダール指数・集中度指数）による論文数・引用数の集中化の動向

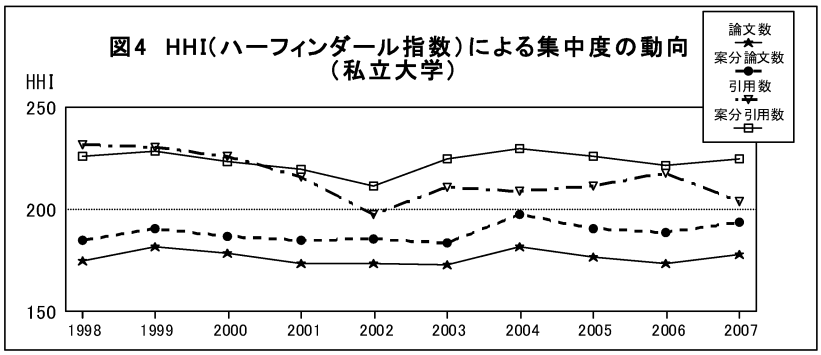
		1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
国立大86校	論文数	430	435	434	429	436	438	449	444	445	453
	案分論文数	415	419	415	411	418	421	433	432	440	442
	引用数	579	559	605	567	594	577	580	569	579	594
	案分引用数	565	552	575	552	569	556	574	576	574	603
私立大502校	論文数	175	182	179	174	174	173	182	177	174	178
	案分論文数	185	191	187	185	186	184	198	191	189	194
	引用数	232	231	226	216	198	211	209	212	218	204
	案分引用数	226	229	224	220	212	225	230	226	222	225
国公私大全体665校	論文数	224	230	228	224	227	226	235	229	228	230
	案分論文数	222	226	223	220	223	223	232	227	230	230
	引用数	323	314	341	311	319	314	318	314	320	323
	案分引用数	324	320	332	314	323	314	327	325	325	344

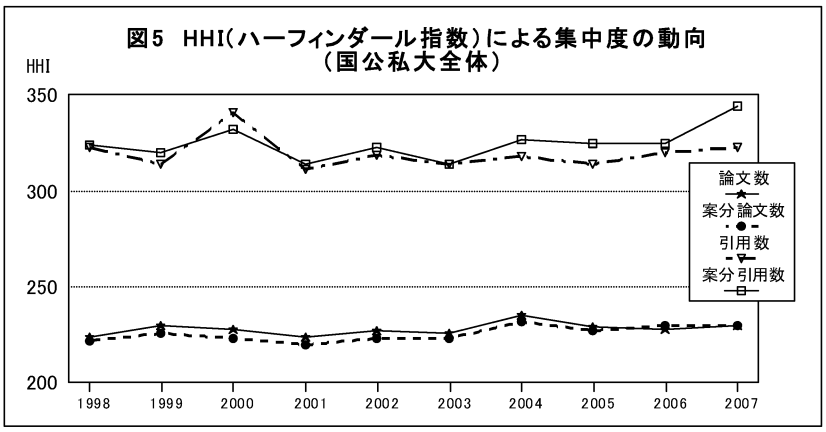
ここでは国立大学，私立大学，さらに国公私合わせた全大学についてそれぞれHHIを計算した<表 8>.

国立大学—国立大学の論文数のHHIは2003年以前も多少とも上昇傾向にあったが，2003年から2004年にかけて438から449へと相対的に大きな上昇幅となり，以後この程度の値が維持されている.案分論文数でも2004年に上昇がみられ，これはその後さらに上昇しつつあるとみてよかろう<図 3>.案分値のHHIの方が単純集計値より低いのは，案分により論文数の分布がより平準化されるということである.

引用数ではその集中度が論文数より格段大きい点にまず注目するべきであるが，振れ幅が大きく趨勢は読み取り難い．それでも案分引用数では2004年以降，HHIの上昇傾向がみてとれる.とくに単純集計値と案分値の差が2004年以降は消失し,2007年には逆転して案分引用数のHHIの方が大きく出ている．すなわち高引用論文の大規模大学への集中という，先にみた傾向がここでも現れている．総じて，国立大学内での論文，特に高引用度論文の2004年以降の集中化傾向はHHIからも明らかであろう.







私立大学—私立大学の場合，大学数が多く HHI は国立大学に比べて低い値になる<**図 4**>．論文数については一定の傾向は読み取れないが，それでも案分論文数をみると，2003 年から 2004 年にかけて一段の上昇があるようにみえる．ここで仮説としては，法人化に伴う国立大学での論文の集中化が，共著関係を通じて私大論文の集中化に波及しているのではないか，すなわち具体的には，大規模国立大と有力私大間の共著論文が相対的に増加してきているのではないかということが考えられる．引用数においては 2002 年まで集中度は低落の基調，すなわち平準化の趨勢にあったが，2003 年以降反転傾向にある．ここ 10 年ほど，短大の 4 年制転換等で私大数が増加しているから HHI が低下するのは当然であるが，その中にあって私立大学でも，高引用度論文の有力私大への集中化がここ数年進んでいるのではなかろうか．

大学全体—国公私大全体についての HHI は，国立大学における集中化傾向はよほど希釈され，全体的集中化は明確ではない<**図 5**>．すなわち，私立短大の 4 年制化の他，公立大でも看護学校の大学化等も含めて，総体的に大学数が増加しており（<**表 1**>標榜掲載大学数欄参照），これが国立大学での集中化を打ち消す方向に働いている．それでも，案分引用数では，2004 年以降の HHI 上昇をみてとることができ，高引用度論文の集中化がこ

こに現れているといえよう.

5 おわりに

国立大学の 2004-2009 年度にわたる 6 年間の第 1 期中期計画，中期目標達成度の「成績表」が，2009 年 3 月 26 日に国立大学評価委員会から公表され，新聞紙上で話題になった．評価結果の良否は次期中期計画期間における運営費交付金の増減に反映されるとのことである．格差こそ活力の源泉論，適者生存主義，比較優位説等はそれとして，当座の弱者にいたずらに市場からの退場を促すばかりになってしまっては，肝心の競争環境，市場の維持保全がおぼつかない．ここに，強者助長と弱者援護との適切な政策配合，ポートフォリオが必要であることは明らかであろう．そのための適確な現状把握と動向分析のために，各種の数量的指標の開発の必要性が一層高まってきていると思われる．本稿においてその一端を示すことができたとすれば幸いである.

参考文献

[1] 杉森純:地方大－揺らぐ研究基盤,読売新聞，2008 年 11 月 2 日朝刊 17 面

[2] Thomson Reuters: National Citation Report (NCR)
http://www.thomsonreuters.com/products_services/scientific/National_Citation_Report

[3] 内閣府（科学技術政策・イノベーション担当）:独立行政法人，国立大学法人等の科学技術関係活動に関する調査結果（平成 19 事業年度）平成 20 年 10 月 31 日，671p. 2008
http://www8.cao.go.jp/cstp/siryo/haihu77/haihu-si77.html

[4]根岸正光：ISI データベースにおける発表論文数，被引用回数に基づく『大学ランキング』作成の方法と問題点，情報知識学会誌,Vol.16,No.2,p.13-18,2006.

[5]根岸正光:ISI データベースにおける論文の『引用力』による交付金・補助金の大学別配分シミュレーション，情報知識学会誌,Vol.18,No.2,p.131-142,2008.

[6] 根岸正光:論文引用度指数ランキング－納税者の視点で資金効率を評価，大学ランキング 2010 年版，朝日新聞社，p.222-231，2009.5 刊行予定.

[7] 根岸正光: ISI・論文引用ランキング－『引用力』による配分額を初試算，大学ランキング 2009 年版，朝日新聞社，p.216-225，2008，ISBN 978-4-02-274520-0.

[8]公正取引委員会:企業結合審査に関する独占禁止法の運用指針，40p. 2007.3.28
http://www.jftc.go.jp/pressrelease/07.march/07032801-01-tenpu03.pdf

ネットワーク指標を用いた学際的な論文の抽出

Extraction of Interdisciplinary Papers by Network Indicators

梶川裕矢，森純一郎
Yuya KAJIKAWA, Junichiro MORI

東京大学大学院工学系研究科
Graduate School of Engineering, The University of Tokyo
〒113-8656 東京都文京区弥生2-11-16工学部9号館
E-mail: takeda@biz-model.t.u-tokyo.ac.jp

我々が直面している持続可能性,少子高齢化といった複雑な課題は単一の学問領域では解決できず,領域間に跨った学際的な研究が重要であるとされている.本研究では,引用ネットワーク分析を用いて,そのような学際的な論文を抽出する手法を提案する.学際性を測定するネットワーク指標を用いて,持続可能性学をケースとして分析を行った結果,複数の指標の組み合わせにより,学際的な論文を効率的に抽出可能であることが分かった.

Clustering using cocitation and bibliographic coupling is an effective tool to analyze the structure of scientific research, but its arbitrariness in the setting of clustering threshold is a problem. This study tracked modularity of citation networks in research domains, and we found that there are three stages in clustering of citation networks and it is universal across our case studies. In the first stage, core clusters in the domain are formed. In the second stage, peripheral clusters are formed, while core clusters continue to grow. In the third stage, core clusters grow again. By focusing on elementary process in clustering, we can understand a hierarchical structure of domain and judge the location of each paper, which cannot be seen in the final clustering results.

キーワード: 引用分析，ネットワーク分析，学際性，分野横断性
citation analysis, network analysis, multidiscipline, interdiscipline

1 緒言

引用ネットワーク分析は,日々膨大な数が出版される学術論文の全体の構造を俯瞰し,目的に応じて適切な論文を抽出するための有効な手法である.引用ネットワーク分析においては,共引用,書誌結合,直接引用といった手法を用いてリンクを形成したのち,被引用数などの統計的集計量の算出,ネットワークの可視化やクラスタリングが行われる[1,2].また,時系列分析や論文の平均出版年の算出により萌芽的な研究領域を検出することが可能である[3,4].

それらの分析により,複雑な学術研究の全体像を粗視的に把握することが可能

である.しかし,膨大な量の情報を単純化することによる情報の見落としは避けられない.従って,利用者の目的に応じた情報の縮約の方法を構築する必要がある.

本研究の目的は,学際的な論文の抽出を行うための手法を構築することである.本研究では学際性を測定する複数の指標を設定し,その妥当性と有効性を評価する.分析の対象としてはサステイナビリティ学を取り上げる.持続可能性や少子高齢化といった複雑な課題は単一の学問領域では解決できず,領域間に跨った学際的な研究が重要であるとされている[5,6].Kajikawaらは引用分析を用いて,サステイナビリティ学が,農業,森林,水,エネルギーなど持続可能とすべき対象によって分かれていることを明らかにしている[5].

しかし,それらは本来相互に関連している.例えば,森林の存在が農業の生産性に影響を与えるし,造水にはエネルギーが必要だ.本研究では,近年,計量書誌学の分野で研究が盛んになってきた学際性を測る研究[7,8]において用いられている複数の指標を,サステイナビリティ学に対して適用し評価したので報告する.

2 手法

媒介中心性はネットワークの中で他のノードを結びつける機能を測定する指標としてよく用いられる.Leydesdorff はこの媒介中心性を引用ネットワーク分析に適用し,媒介中心性(Betweens Centrality, B)を用いて学際性を評価できるとしている[7].また,Rafols は学際性を測る指標として,当該文献が引用している文献の引用文献(RefRef)が属している,カテゴリー数(N),Shannon の情報量(H),Simpson の多様性(I),Striling(S)を用いて評価している.各指標の定義は以下の通りである.

$$H = -\sum_i p_i \ln p_i$$

$$I = 1 - \sum_i p_i^2$$

$$S = 1 - \sum_{i,j} s_{i,j} p_i p_j$$

ただし,p_i はカテゴリーi に属する RefRef の割合,$s_{i,j}$ はカテゴリーi, j 間の距離である.

我々のサステイナビリティ学分野における既存研究では,引用ネットワークを Newman 法[9]法を用いてクラスタリングを行い,可視化することで全体像の分析を行っている.従って,本研究でも同様に Newman 法を用いたクラスタリングを行い,各クラスタをカテゴリとした.また,$s_{i,j}$ としては,クラスタ間のリンク数を各クラスタのノード数で正規化したものを用いた.

上記指標に加え,クラスター分散度(Participation coefficient, P)[10], 次数中心性(Degree, D)を測定した.

3 結果と考察

表 1 に各指標間のPearsonの相関係数を示す. Rafolsの提案している,N, H, I, S はどれも互いに相関性が高い.Pは他の指標との相関性が低く,Leydesdorffの提案しているBはDとの相関性が高い.これは,Rafolの指標群はどれも属しているクラスターの数, すなわち,Nの影響を強く受けるためと考えられる.また,Bはしばしばネットワークにおける媒介性の指標

として広く使われるが,本質的には中心性の指標であり,リンク数が多いほど中心性も高くなるためBとDの相関性が高くなったといえる.一般に引用ネットワークのクラスタリングにより得られた各クラスタの特徴は各クラスタにおいてハブとなっている論文の影響を色濃く受ける.従って,媒介中心性を学際性の指標とすることは少なくとも今回の場合適切でないように思われる.

表1 各指標間の相関性

	N	H	I	S	P	B	D
N	1	0.89	0.82	0.82	0.37	0.22	0.29
H		1	0.99	0.99	0.43	0.16	0.2
I			1	1	0.42	0.14	0.18
S				1	0.42	0.14	0.18
P					1	0.21	0.22
B						1	0.74
D							1

そこで,NとPに対し,どちらの方が学際性の高い論文を抽出する精度が高いか評価した.評価のためにまず,各クラスタの特徴付けを行った.特徴付けには,そのクラスタに含まれる論文が出版されているジャーナル,そのクラスタのハブ論文のタイトルとアブストラクトの情報を用い著者が行った.NとPの値が高いそれぞれ２０論文を,A:クラスタの特徴と一致,B:他クラスタの特徴と一致,C:学際性が高い,という３つのカテゴリに分類した.
結果,Nに対しては,A=5, B=9, C=6,Pに対しては,A=1, B=9, C=10,という結果が得られた.すなわち,少なくとも今回のサステイナビリティ学に対しては学際性の指標としてクラスタ分散度が有効であることが分かった.

4 結言

引用ネットワークのクラスタリングや可視化は学術分野の全体像を把握するための有効な手法であるが,しばしば萌芽的な分野,学際性の高い分野の存在を見落としがちである.本研究では引用ネットワーク中の各論文のトポロジカルな位置を表す複数の指標をサステイナビリティ学における論文集合に対して適用した.各指標間の相関性を分析することで,各指標を３つのカテゴリに整理した.これらのうち,媒介中心性はリンク数との相関が強く学際性の測定には不適切であることが示唆された.属するクラスタ数とクラスタ分散度の比較では, クラスタ分散度の方が学際性の評価指標として優れていることが示唆された.

参考文献

[1] Small, H.: “Visualizing science by citation mapping”, Journal of the American Society for Information Science, Vol.50, pp.799-813, 1999.

[2] Börner, K.; Chen, C.; Boyack, K.W.: “Visualizing knowledge domains”, Annual Review of Information Science & Technology, Vol.37, pp.179–255, 2003.

[3] Small, H.: “Tracking and predicting growth areas in science”, Scientometrics, Vol.68, pp.595–610, 2006.

[4] Shibata, N.: Kajikawa, Y.: Takeda, Y.: Matsushima, K.: "Detecting emerging research fronts based on topological measures in citation networks of scientific publications" Technovation, vol.28,

pp.758-775, 2008.
[5] Kajikawa, Y.: Ohno, J.: Takeda, Y.: Matsushima, K.: Komiyama, H.: "Creating an academic landscape of sustainability science: an analysis of the citation network", Sustainability Science, Vol.2, pp.221-231, 2007.
[6] Kajikawa, Y.: "Research core and framework of sustainability science", Sustainability Science, Vol.3, pp.215-239, 2008.
[7] Rafols, I., Meyer, M.: “Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience”, Scientometrics, in press.
[8] Leydesdorff, L.; “Betweenness Centrality as an Indicator of the Interdisciplinarity of Scientific Journals”, Journal of The American Society For Information Science and Technology, Vol.58, pp.1303–1319, 2007.
[9] Newman, M.E.J.: “Fast algorithm for detecting community structure in networks”, Physical Review E, VOl.69, artNo.066133, 2004.
[10] Guimera, R.: Amaral, L.A.N.: “Functional cartography of complex metabolic networks”. Nature, Vol.433, pp.895-900, 2005.

産学連携に関するWeb情報の分析　－大学サイトの事例研究－

University-Industry Collaborations in Academic Websites: A Case Study of Seven National Universities

柿沼澄男*，孫媛，西澤正己，大山敬三，根岸正光
Sumio KAKINUMA, Yuan SUN, Masaki NISHIZAWA, Keizo OYAMA, Masamitsu NEGISHI

国立情報学研究所
National Institute of Informatics
〒101-8430　東京都千代田区一ツ橋２－１－２
E-mail: kakinuma@nii.ac.jp

産学連携の状況は，これまで共著論文，特許引用，産学連携統計により示されてきた．筆者らは，これらの指標とは別に，大学のWebサイトに現れた情報により，産学連携の状況を分析・計量化することに取り組んでいる．本報告においては，７大学（北海道大学，東北大学，東京大学，名古屋大学，京都大学，大阪大学，九州大学）のWebサイトに現れた産学連携に関する情報を整理分析した結果，明らかになった特徴を示す．

Collaborations in scientific research between universities and industries have been studied mainly based on coauthorship of academic publications, citations of patents, or related statistics. The authors intend to analyze and measure the relations of universities with industries through the information on universities' web pages. In this study, we carried out case studies of seven major national universities to show the nature of information on the university-industry collaborations released on universities' websites in Japan.

キーワード：Web計量学，産学連携，共同研究，大学のWebサイト
Keyword: webometrics, university-industry collaboration, research collaboration, academic website

１　はじめに

近年，インターネット上の情報が爆発的に増大する中，各大学もWebサイトの拡充に務め，その情報量は増大の一途を

たどっている．産学連携の状況は，これまで共著論文，特許引用，産学連携統計により示されてきたが，このようなインターネットの状況を考慮すると，Web 上の情報を何らかの方法で分析することが必要であると考えた．

欧米では，産学連携についてWebデータを収集し，相互のリンク関係，同時リンク関係を用いた分析手法の有効性が示されている［1］［2］．しかし，筆者らが日本の状況を試みに調査・分析したところ，こうした方法が単純には適用できないことが明らかになった．

そこで，産学連携活動の多い大学を対象として，Web データの特徴・構造について調査し，それを踏まえた上で，日本の産学連携の実態を反映するWebデータの収集方法や適切な指標を検討・開発することにした．本報告では，産学連携活動が盛んな7大学（北海道大学，東北大学，東京大学，名古屋大学，京都大学，大阪大学，九州大学）を対象として，Web サイトに現れた産学連携に関する情報を整理分析した結果，明らかになった特徴を示す．Web データの収集方法や適切な指標を検討・開発するのは，今後の課

表1　産学連携関連組織

	組織構造		
	全学		関連外部組織
	知的財産本部名称	本部内組織	
北海道大学	知財・産学連携本部	連携推進部 　連携企画室 　包括連携室 知財財産部 事業化推進部 東京支部 【学内関連組織】 創成科学共同研究機構 （超学問領域の創成とその 研究成果を世界へ向けて発信）	・北海道ティー・エル・オー（株） http://www.h-tlo.co.jp/www-home/ ・北大ビジネススプリング （大学発ベンチャーの支援：中小企業基盤整備機構運営） http://www.ho-bis.jp/
東北大学	産学連携推進本部	研究推進部 　連携企画室 　国際連携室 知的財産部 事業化推進部 　事業化支援室 　地域連携室 高度人材キャリアセンター ナノテク支援センター	㈱東北テクノアーチ http://www.t-technoarch.co.jp/
東京大学	産学連携本部	産業連携協議会 産業連携研究推進部 　Proprius21（共同研究新スキーム） 　TLF室（産学連携人材育成） 知的財産部（知的財産の管理と活用） 事業化推進部（大学発ﾍﾞﾝﾁｬｰ支援・起業教育） 産学連携Gr（業務支援）	・㈱東京大学TLO（CASTI） http://www.casti.co.jp/ ・㈱東京大学エッジキャピタル（UTEC） http://www.ut-ec.co.jp/
名古屋大学	産学官連携推進本部	産学官連携推進室 　国際連携部 　連携推進部 　知的財産部 　起業推進部	（財）名古屋産業科学研究所（中部TLO） http://www.nisri.jp/

表1　つづき

	組織構造		
	全学		関連外部組織
	知的財産本部名称	本部内組織	
京都大学	産官学連携本部	産官学連携センター 　国際連携推進室 　産官学連携推進室 　知的財産室 　ベンチャー支援室開発室（京大ﾍﾞﾝﾁｬｰﾌｧﾝﾄﾞと連携） 　法務室（法的実務支援） 　分野拠点（メディカル・バイオ分野、理工農学分野、ソフトウェア・コンテンツ分野） 【学内関連組織】 京都大学「医学領域」産学連携推進機構（KUMBL） 　産学連携オフィス 　インキュベーション・プラザ ベンチャー・ビジネス・ラボラトリー	関西ティー・エル・オー㈱ http://www.kansai-tlo.co.jp/index.php 日本ベンチャーキャピタル㈱ http://www.nvcc.co.jp/
大阪大学	産学連携推進本部	イノベーション創出部 知的財産部 総合企画部 事務部門（研究推進部）	・NPOおおさか大学企業支援機構 http://www.aiue-osaka.org/index.html ・阪大イノベーションファンド http://www.crcast.osaka-u.ac.jp/innovation_fund/ ・NPO青い銀杏の会 http://www.osaka-u.com/ ・大阪TLO http://tlo.casi.osaka-u.ac.jp/
九州大学	産学連携推進機構	知的財産本部（1センター＋7部門） 　国際産学官連携センター 　企画部門 　リエゾン部門 　プロジェクト支援部門 　技術移転部門 　起業支援部門 　デザイン総合部門 　　アジアDLO 　事務部門 産学連携センター 　リエゾン部門 　デザイン総合部門 　プロジェクト部門 　連携部門 ベンチャー・ビジネス・ラボラトリー	㈱産官学連携九州機構（九大TLO） http://www.k-uip.co.jp/

題である．

なお，この調査データは平成２１年３月時点の各大学の Web サイト上のデータある［3］．

2　産学連携関連組織

平成１５年度から実施されている，文部科学省の「大学知的財産本部整備事業」により，各大学には知的財産の創出・取得・管理・活用を戦略的に実施するための組織である知的財産本部（産学連携窓口）が整備されている．大学における産学連携活動を把握するためには，この組織から発信される情報を分析することが重要であると考え，情報収集の際に参考とするため，Web サイト上に掲載されている組織構造を調査した．

表１は，７大学の知的財産本部の組織構造と関連外部組織を示したものである．

表2　産学連携の成果

			北大	東北大	東大	名大	京大	阪大	九大
成果	共同研究	実施件数	○	○	○	○	-	○	○
	受託研究	実施件数	○	○	○	○	-	○	○
	ベンチャー	起業数	○	○	○	○	○	-	○
	寄付講座		○	○	○	○	○	○	○
	知的財産関連収益		○	-	○	○	○	○	-
	TLO	発明届出数	○	○	○	○	○	○	○
		特許出願数	○	○	○	○	○	○	○
		移転等数	○	-	○	○	○	○	○

3　産学連携の成果

産学連携の成果を示す指標として，共同研究や受託研究の実施件数などがあるが，これらの情報がどの位掲載されているかは，Web サイトがファクトデータを収集する手段となり得るかを判断する上で重要となる．

表２は，Web サイト上に公開されている各大学の産学連携成果である．この種類のデータは，様々な評価活動で求められる情報であることから，各大学とも Web サイト上に掲載している．

掲載されている数値を，文部科学省から公表されているものと比較すると，数値が異なっている大学があり，ファクトデータとして使用する上で問題がある．

4　研究成果のプレスリリース等

大学内の産学連携関連情報は，様々な Web サイト（個別研究紹介，研究科・研究室紹介，プレスリリース等，年報類，知的財産本部，TLO など）に掲載されている．これらの情報は膨大であり，最初から全ての Web サイトを対象として分析することは非効率になる恐れがあるため，最初は特定の種類の Web サイトに絞って分析を進めていくことが必要である．筆者らは，まず分析を進める対象として，プレスリリース等を取り上げることにした．

近年，広報の一環として，研究成果の内容について，その概要をプレスリリース等（呼称は大学によって様々である）として Web サイトに掲載している大学が多くなっている．ここには，報道機関へ配布された資料や記者発表資料が，そのまま掲載される場合もある．プレスリリース等に掲載される情報は，その大学が重要と判断したものであり，研究の分野，相手企業，日付など多くの情報を含んでいるため，分析の対象としては格好の材料である．

表３は，７大学のプレスリリース等の Web サイトにおける構造を示したものである．北海道大学と大阪大学を除く５大

表3 研究成果のプレスリリース等

	北大	東北大	東大		名大	京大	阪大	九大
階層1	×	Topページ	Topページ	Topページ	Topページ	Topページ	×	Topページ
階層2	×	新着情報 （受賞・成果等）	Information （広報・情報公開）	学術情報 （一覧）	お知らせ／イベント （研究教育成果情報）	研究	×	プレスリリース・広報誌
階層3	×	研究成果	記者発表	標題・概要一覧	標題一覧	研究成果	×	プレスリリース
階層4	×	年度別標題一覧	研究成果	本文	本文	年度別標題一覧	×	年度別標題一覧
階層5	×	概要	標題一覧			本文	×	本文
階層6	×	本文	本文				×	
収録期間	×	2002年から	2004年から		2001年から	2004年から	×	2006年から

学では，プレスリリース等がWebサイトに掲載されている．Web サイト上における構造（「年度毎／通史」，「本文のリンク先が学外にも及んでいるか」など），掲載内容（「研究成果のみ／研究成果以外も含む」など）は，大学ごとに異なっている．そのため，情報をクロールして収集する方法を開発する場合には，これらのことを考慮する必要がある．

5 おわりに

筆者らはWebサイトに現れた産学連携情報の計量化を目指しているが，本報告はその第一歩である．今回の調査により，いくつかの大学のWebサイト構造を確認することができたので，今後は，対象大学を拡大して調査することと，実際にWeb サイトをクロールして，データを収集・分析することを予定している．

参考文献

[1] Vaughan, L., You, J.: “Mapping business competitive positions using Web co-link analysis”, In: P. Ingwersen, B. Larsen (Eds), Proceedings of ISSI 2005 - the 10th International Conference of the International Society for Scientometrics and Informetrics, pp.534-543, Stockholm, Sweden, July 24-28, 2005.

[2] Vaughan, L.: “Visualizing linguistic and cultural differences using Web co-link data”, Journal of the American Society for Information Science and Technology, Vol.57, No.9, pp.1178-1193, 2006.

[3] 各大学の URL

北海道大学

http://www.hokudai.ac.jp/index.html

東北大学

http://www.tohoku.ac.jp/japanese/

東京大学

http://www.u-tokyo.ac.jp/index_j.html

名古屋大学

http://www.nagoya-u.ac.jp/

京都大学

http://www.kyoto-u.ac.jp/ja

大阪大学

http://www.osaka-u.ac.jp/index.html

九州大学

http://www.kyushu-u.ac.jp/index.php

論文からの解析手順知識の網羅的抽出から分かる解析技術の経年的変化の分析

Interannual Change of Bioinformatics Analysis Techniques by Workflow Extraction from Biological Papers

荒木次郎

Jiro ARAKI

＊(株)三菱総合研究所 先端科学イノベーショングループ

Mitsubishi Research Institute, Inc.　Advanced Science and Innovation Group

〒100-8141 東京都千代田区大手町 2-3-6

E-mail: jiro@mri.co.jp

我々は，バイオインフォマティクス解析のためのセマンティックウェブサービスの実現を目指して，解析ウェブサービスの充実とともに，学術論文中から解析手順知識(ワークフロー)を半自動抽出する技術を開発し網羅的な収集を試みている．このようなワークフローを利用することで，ユーザの解析目的に対してそれを実現するための解析方法の組合せを提示するためのサービスを提供する．本稿では，網羅的な解析手順知識の収集によって初めて分かる解析技術の経年的変化の分析結果を中心に報告する．

We investigate how to collect and reuse workflows (= knowledge of analysis techniques) in order to develop a semantic web service system for bioinformatics. We have developed a semi-automatic workflow extraction method from biological papers. Our goal is to develop a service to provide workflows for biologists' analysis purposes. In this paper we report interannual changes of bioinformatics analysis techniques, which are found by analyzing collected workflows.

キーワード：セマンティックウェブ，ウェブサービス，ワークフロー，バイオインフォマティクス，知識抽出

Semantic Web, Web Service, Workflow, Bioinformatics, Knowledge Extraction

1　はじめに

近年バイオサイエンス分野では，ゲノム解読やプロテオーム解析に代表される多くの網羅的解析プロジェクトが進められ，莫大なデータが生産され続けている．しかし，それらのデータはあくまでも生命の断片情報であるため，バイオインフォマティクス手法を利用して断片知識の統合や大規模データからの知識発見を行い，生命を理解しようとしている．

例えば，2001 年に国際ヒトゲノムコンソーシアムによってヒトゲノム解読の成果が雑誌 Nature で発表された[1]．このプロジェクトでは，まず 30 億文字にも及ぶヒトゲノム配列データを各国の研究機関が分担して解読した．断片状のゲノム配列データは繋ぎ合せるための解析が行われ，次にゲノム配列全体の中から遺伝子に当る部分を見つけ出すための分析が行われた．最後にそれぞれの遺伝子がどのような機能をもつタンパク質の設計図であるかを予測するための解析が行われた．これらのバイオインフォマティクス解析も各国の研究機関によって分担して行われ，その解析手順や結果が 60 ページに及ぶ論文の中にまとめられている．

我々は論文中の記述からこのような解析手順をワークフローの形で抜き出す作業を行った[2]．図 1 は，ヒトゲノム解読論文から抜き出したワークフローである．図内の長方形が解析ツールを表し，その前後に繋がる楕円形が解析ツールの入出力データ，解析に利用されたデータベースを表している．図から分かるように，1 生物種のゲノム配列を解析するだけでもツールやデータベースの非常に複雑な組合せが必要であり，ワークフローは各国の研究者が持ち寄った知識の集積と言える．

2001 年のヒトゲノム解読以降，現在までに数百の生物種のゲノム解読が完了し，論文が公開されている．それぞれの論文では，過去の論文の解析手法を参考にしつつも，その生物種や研究目的に特化した手法にアレンジしていると思われる．また，時代とともにより複雑で高度な解析が行われていると思われる．そのためバイオ研究者が研究を進めていく為には，解析ツールやデータベースの個別の知識とともに，このような解析手順の知識が必要である．

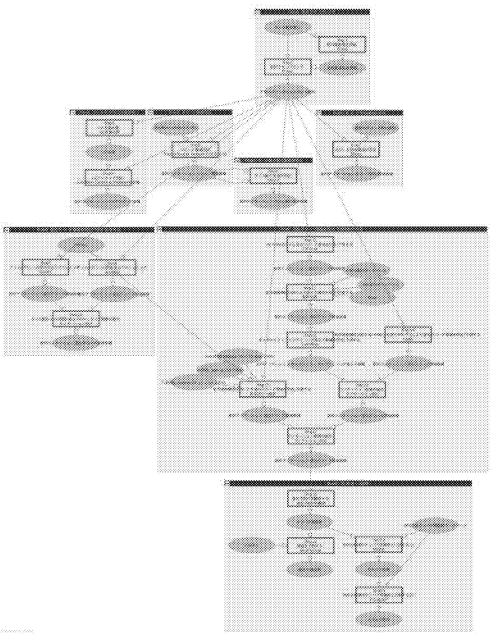


図 1 論文から抽出した解析手順知識の例[2]

そこで我々はバイオサイエンス系論文からバイオインフォマティクスの解析手順知識を抽出して「ワークフロー事例データベース」として蓄積し，バイオ研究

者の目的に合った解析手法を提示するための「ワークフロー検索システム」の開発を進めている(図 2参照)．その中で我々はまずワークフローやその構成要素である解析ツールやデータベースを統一的に表現するための「ワークフローオントロジー」を作成した[2]．そして論文から網羅的にワークフローを抽出するための「ワークフロー自動抽出アルゴリズム」の初期版を開発した[3]．本手法では，係り受け解析を利用し，さらにオントロジーによってその組み合わせの制約を規定することで抽出精度を上げている．

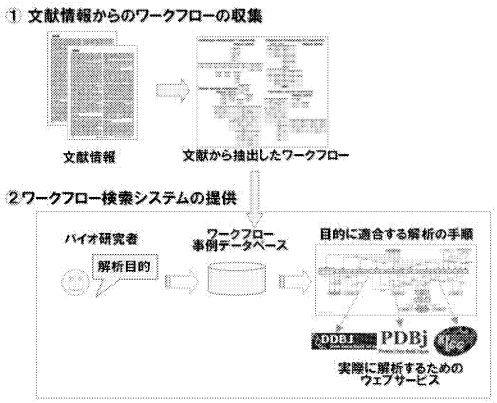


図 2 解析手順知識の収集と活用の流れ

我々が論文からワークフローを収集する主目的はセマンティックウェブサービスの実現であるが，付加的な目的として文献情報による研究分野俯瞰への貢献がある．従来の研究では主に分野キーワードによる分析が行われてきたが[4]，研究手法や研究成果の応用などを知るためにはキーワード間の関係性も考慮したより詳細な分析が必要となってくる．論文中から研究手法を抽出する我々の手法がこの課題にどれくらい貢献できるかを分析するのが本稿の目的である．

本稿では，まず2章，3章において既往研究で開発した解析手順知識のデータ表現方法[2]と，論文からの自動抽出手法[3]についてそれぞれ説明する．次に4章において解析手順知識の網羅的収集から分かる解析技術の経年的変化の分析結果についてまとめる．

2　解析手順知識のデータ表現手法

本章ではまずバイオサイエンス系論文に記述されているバイオインフォマティクスの解析手順知識の構成内容を定義する．次にそれをデータベース化し，検索等で再利用するためのデータ表現方法について説明する．

2.1 解析手順知識の定義

本研究において論文記述中から抽出する解析手順知識とは，バイオサイエンス系論文に記述されたバイオインフォマティクス解析の手順である．通常，Methodsなどの名前の章が設けられ，解析手順や利用材料などが記述されている．

解析手順の知識は自然文で記述されているが，その構成は図 3 に示す要素から成る．「解析ツール」は解析対象となる「入力データ」を入力とし，解析結果である「出力データ」を出力する．また検索を伴う解析ツールの場合は「データベース」が入力される．また，解析ツールには「分類する」「検索する」などの「解析処理機能」が付与される．このように解析ツール単独では非常にプリミティブな機能しか持たないが，特定種の入力データやデータベースを伴うことで生物学的な「解析目的」を達成することができる．

これらの要素からなる基本セットが組

み合わせられることで複雑な解析ワークフローが実現する．例えば，ある解析ツールの出力データが別の解析ツールの入力データとなる場合(直接処理)や，入力データが複数の解析ツールの入力データとなる場合(並列処理)などが考えられる．

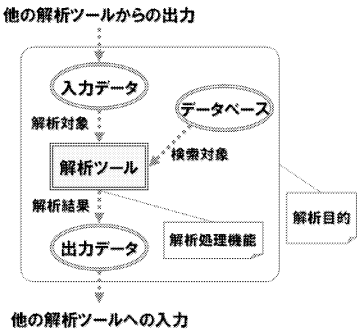


図 3 解析手順知識の定義

2.2 データ表現方法

論文から抽出したワークフローを表現するためのオントロジーを図 4，図 5 に示す．ワークフローは解析目的(Purpse)単位で階層的に構成関係(composedOf)が表現され，最終的には解析ツール(Tool)と入出力データ(Data)，データベース(Database)の組み合わせで表される．解析ツールに付与される解析処理機能(Function)は「予測する」などのような，解析目的よりもよりプリミティブな機能である．

本オントロジーの特徴は，これらのワークフロー構成要素が構成関係上のつながりだけでなく意味的なつながりを持たせている点である．即ちそれぞれの構成要素は生物学ドメインオントロジー(図 5)を介して生物学的な意味付けがなされ，推論機能によって既存のワークフロー以外のワークフローを新たに作り出すことが可能となる．

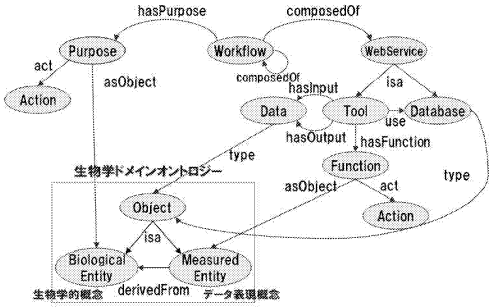


図 4 ワークフローオントロジー[2]

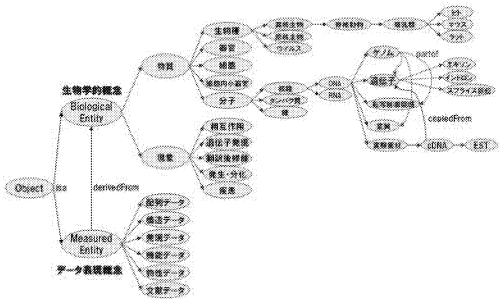


図 5 生物学ドメインオントロジー[2]

3　解析手順知識の自動抽出手法

論文中に自然文で記述された解析手順知識をワークフローの形で網羅的に抽出するために，我々は文単位でのワークフロー自動抽出手法を開発した[3]．手法の概要を図 6 に示す．

本手法は，まず解析ツール名や入出力データ表現名などの固有表現抽出を行い，係り受け解析によって固有表現間の文法的関係を分析する．次にワークフロー抽出のために我々が構築した「格フレーム辞書」を利用して，固有表現間の文法的関係をワークフロー構成関係に変換し不整合がないかをチェックする．

なお，本手法は文単位での抽出に限定されているために，一般的に複数の文に

またがって記述されているワークフローは断片化されて抽出される．そのため，現在は照応関係や暗黙的に記述されていない情報の補完などを手作業で行った上でワークフローを完成させている．

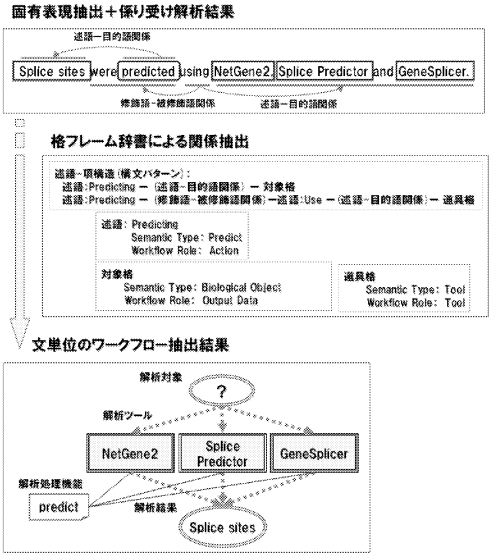


図 6 ワークフロー自動抽出手法[3]の概要

これまでに約400件のゲノム解読関連論文からワークフローを自動抽出し，そのうち 19 件に対して手作業を行い完成させた．

4　解析技術の経年的変化の分析

本章では収集したワークフローデータを利用して，バイオインフォマティクスの解析技術がどのような目的で，どのような組み合わせで利用されているかを分析する．

4.1 利用したデータ

分析に利用したワークフローデータは，ゲノム解読関連論文から自動抽出した後に手作業で完成させた 2000～2005 年発表の 19 件のワークフローである．

4.2 分析結果

図 7 にワークフロー規模の経年的変化を示す．ワークフローの規模はそれを構成する解析ツールの数として定義した．図から分かるように，年々構成ツール数は増加しており，ワークフローはより複雑化している．

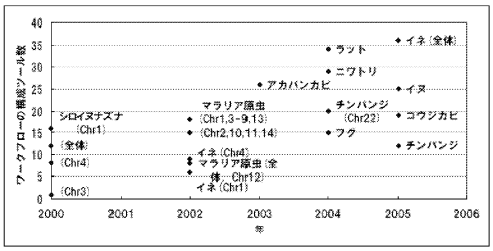


図 7 ワークフロー規模の経年的変化

次にワークフローがどのような解析の目的で構築されているかを見る．図 8 は全 19 件のワークフローを発表年順に縦に並べ，それが 5 つに分類した解析目的のいずれに該当するかを示した．

いずれのワークフローもゲノム解読に関連する論文から抽出したものなので，ゲノム上にどのような遺伝子がコードされているかを分析するアノテーションが解析目的に含まれている．また，ゲノム上の遺伝子セットからその機能を調べる機能解析も従来から多く行われている．一方，この 6 年間の中での特徴的な経年変化も見られる．ゲノム解読が盛んに行われるようになった数年後の 2004 年以降，急に比較ゲノム解析が行われるようになっている．これはそれまでにさまざまなゲノム配列が決定されデータが蓄積されることで，生物種間の比較分析が可能となったためと考えられる．

なお，ここでは示していないが，利用している解析ツールの経年的変化もある．まだデータがあまり蓄積されていなかっ

た 2000 年頃は既知データを利用しない Ab initio な遺伝子領域予測ツール(GenScan など)が多く用いられていた．一方，それ以降はデータベースの充実により，既存データとの比較によって遺伝子を予測する手法がとられるようになった．

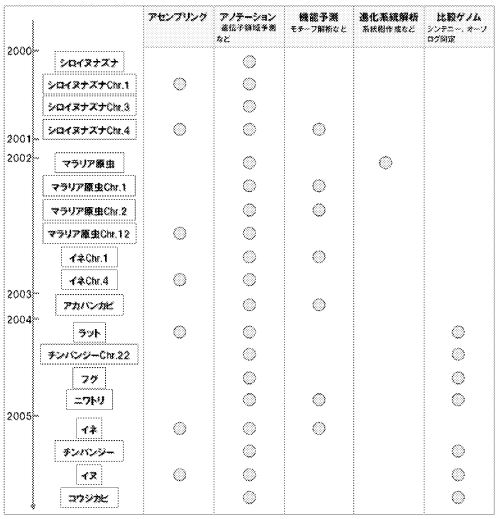


図 8 ワークフローの解析目的の経年的変化

5　まとめ

本稿では，バイオサイエンス系論文からバイオインフォマティクス解析のワークフローを抽出するための手法を紹介した．そして抽出したワークフローデータを利用することで解析技術の経年的変化を分析できることを示した．

従来のビブリオメトリックスでは分野キーワードや引用関係などをもとに研究動向や関係性を見ることが主であったが，本研究では文献を研究手法という新たな視点で整理することを提案した．このような視点で整理することは，新たな手法によって急激に研究が進む可能性のあるバイオ分野などで非常に有効であると考えられる．

今後は，ワークフローの自動抽出手法の高度化とともに，さまざまな分野への適用を考えている．

参考文献

[1] Lander, E.S et.al:“Initial sequencing and analysis of the human genome”, Nature, Vol.409, No.6822, pp. 860- 921, 2001.

[2] 荒木次郎, 川本祥子, 藤山秋佐夫, 菅原秀明, 大久保公策, 武田英明:「文献からのバイオサイエンス研究手法の収集・整理による研究支援セマンティック Web サービスの実現」, 第 21 人工知能学会全国大会, 2007.

[3] 荒木次郎, 藤山秋佐夫, 菅原秀明, 武田英明:「バイオサイエンス論文からの解析手法知識自動抽出アルゴリズムの開発」, 第 22 回人工知能学会全国大会, 2008.

[4] 西澤正己, 孫媛 :「キーワード分析による科研費におけるゲノムおよびナノテクノロジー関連研究の動向調査」, 情報知識学会誌, Vol. 17, No. 2, pp 117-122, 2007.

協調フィルタリングを用いた共著関係の予測
Co-authorship Prediction based on collaborative filtering

堀 幸雄† 中山 堯†† 今井慈郎‡
Yukio HORI† Takashi Nakayama†† Yoshiro Imai‡

† 香川大学総合情報センター
Information Technology Center, Kagawa University
〒 760-0016 香川県高松市幸町 2-1
EMail: horiyuki@itc.kagawa-u.ac.jp
†† 神奈川大学理学部情報科学科
Department of Informatin Science, Faculty of Science, Kanagawa University
〒 259-1293 神奈川県平塚市土屋 2946
EMail: nakayama@info.kanagawa-u.ac.jp
‡ 香川大学工学部信頼性情報システム工学科
Department of Information Systems Engineering, Faculty of Engineering, Kagawa University
〒 761-0396 香川県高松市林町 2217-20
EMail: imai@eng.kagawa-u.ac.jp

近年科学技術の専門化が進み，また技術進歩の大規模化等に伴い，多くの研究分野において研究者が共同研究を進める必要性が生じてきた．本研究は論文の共著ネットワークにおける共著予測に内容ベースの協調フィルタリングを適用する手法を提案する．国内論文データベースを用いた評価実験の結果，本手法はリンク構造のみを用いた手法と比較して，予測精度は低いが類似する分野の研究者の共著パターンを知ることができることがわかった．

Recently, various fields of science are becoming increasingly specialization. There is a necessity for performing the research in a lot of fields with collaboration. This paper proposes a recommendation system based on Collaborative Filltering for automatically obtain a research partner from the network. By using data from a bibliographic database, this network is constructed in which the nodes are researchers and two researchers are connected if they have co-authored a paper. The proposed system estimates the usefulness of unknown co-authored researcher using Collaborative Filtering algorithms. This suggested the proposed system can be used for discovering useful researchers.

キーワード:　キーワード: 協調フィルタリング，リンク予測，データマイニング
Keywords: Collaborative Filtering, Link Prediction, Data Mining

1　はじめに

近年，学術分野の専門分化と伴に学術に関する情報料は爆発的に増加し，一人の研究者が研究分野を俯瞰することが困難になってきている．自然科学分野における SCI 収録論文数は年を追うごとに増加しており，研究分野は専門化，細分化が進んでいる．これに対応するために，多くの分野において研究者同士が連携を進め，効果的に新しい研究分野の開拓を行なう必要がある．実際に医療，ロボット，生命情報科学，シミュレーションなどの様々な分野で研究者の連携が必要とされている．このような分野においてはそれぞれの分野の研究者が共同で各々の分野の知識を活用して研究を遂行する．

共同研究の動向は近年大きく変化している．神田らによる調査では SCI 収録論文を共著形態により分類し，単著論文 (著者が一人の論文)，単一機関内共著 (単一の機関に所属する複数の著者による論文)，国内機関間共著 (同一国の複数の機関に所属する著者による論文)，国際共著論文 (異なる国の機関に所属する著者による論文) の増減を報告した [1]．その結果，国内機関間共著と国際共著は 1985 年以降、

ほぼ一貫して増加傾向にあることが示された．研究の遂行において研究機関や国のボーダーを越えた協力関係をより活発に行なうための仕組みが重要であると考えられる．

しかしながら，実際の共同研究は，研究者の個人的な知り合い関係や紹介によって行なわれるのが一般的であり，システマティックに効率的な連携が生まれているとは言いがたい．そこで本稿は論文データベースから研究者の共著ネットワークを抽出し，共同研究しやすい研究者を協調フィルタリングにより推薦するシステムを作成した．情報系の論文データを用いて本システムの共著予測精度を求めた．また本システムと著者間の類似度を用いた方法との比較を行なった．その結果，本システムは共著予測精度は先行研究に劣るものの，研究者にとって近い研究者の共同研究動向を利用するという点において，今後の共同研究の遂行に参考になることがわかった．

2　関連研究

図 1 は神田らによる SCI の一論文当たり平均著者数の推移，図 2 は論文共著形態別の SCI 収録論文数の推移をまとめたものである．これから図 1 からは論文あたりの平均著者人数が単調増加傾向にあること，図 2 からは単著，単一機関内共著による論文が徐々に減少し，機関間共著，国際共著が増加していることがわかる．このような背景には学術論文の出版数の増加や，学問の専門化・細分化のため一人で研究を行うよりも，複数人が集まってそれぞれ得意とする分野の分担を行い，共同研究を行ったほうがより効率的であると考える人が増えたためだと考えられる．

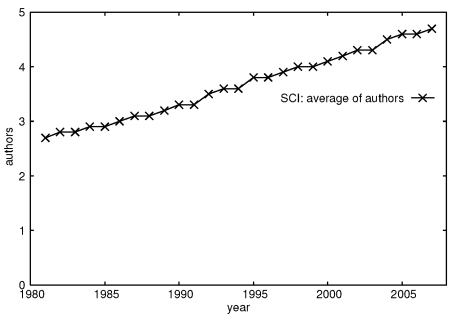


図 1: SCI における 1 論文当たりの平均著者数

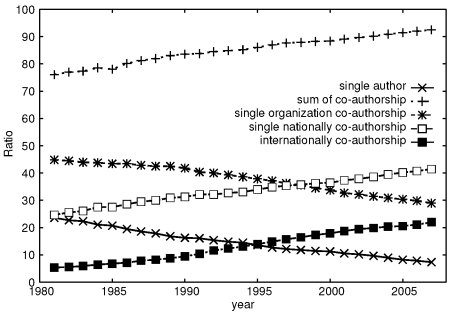


図 2: 共著形態別 SCI 収録論文数の推移

研究における共同研究，共著ネットワークの分析という観点から論文における共著に関するネットワークの指標について分析する研究がある．孫らは日本の学会誌論文を対象とした引用文献索引データベース CJP を用いて研究組織ごとの共著論文の比率から産学連携の現状の分析し，とくに工学の分野で共著の比率が高いことを示している [2]．この研究では共著論文の数を大学と企業の関わりを知るための手がかりとしている．また芳鐘らはこれまで提案された共著ネットワークに関する分析手法をまとめ，それぞれの特徴について考察している [3]．共著ネットワークの研究は共著のパターンを著者と共著頻度をもとに，研究者同士のつながりを調べ，研究者個人の共同研究者選択という振舞いのモデル化を行なっている．この研究では共著頻度や協力係数などの指標の値と共著に関する関係を分析したものであるが，結果を単にネットワーク図を示すだけにとどまり具体的な共同研究者の選択への指針を示すまでには至っていない．

また研究成果の活用，共同研究の推進を目的として研究者の研究内容や活動状況といった情報を Web 上に公開する大学が増えている．廣川らはこの教員データのキーワードの特徴を調べ，研究者に対する検索システム教員マップ作成システムを提案している [6][7][8]．本システムは研究者の入力した活動状況をもとにし，検索要求から連想される研究者をクラスタリングしてユーザに返すものである．また JST の研究開発支援総合ディレクトリ (Directory Database of Research and Development Activities: Read)[5] は国内の研究機関と研究者に関するデータベースであり，国内の研究者に関する最大の規模であると思われる．Read では研究機関，研究者，研究

課題，研究資源の 4 つの情報についてキーワード検索やディレクトリ構造をベースとしたインタフェースで閲覧することが可能である．この情報は産学官連携，人的交流の促進，共同研究の推進を目的として Web を介して誰でも利用することが可能である．しかしながら実際の研究者の連携，共同研究の取り決めは個人的な紹介によって行なわれており，このような研究情報情報を有効に活用されているとは言いがたい．

論文における共著の予測では Nowell らが物理学の論文について著者をノード，共著関係をリンクとしたネットワークから将来の共著関係の予測を行なった [4]．この研究ではネットワーク構造のある部分を元に，残りの部分の構造を予測するものである．具体的には類似した共著関係の持つなどの研究者が将来共著するという仮定のもとに将来の共著関係を予測を行なっている．将来の共著関係を予測する点では本研究と目的は同じであるが，本研究は研究者の所属する組織内の研究者から適した共著相手を推薦し，組織内の共同研究を支援することを目的としている．

またユーザに対して未知のアイテムの評価を予測する協調フィルタリング [16][17][18] に関する研究として様々な応用例が報告されている [13][14][15]．本研究はユーザに過去に共同研究したことをない相手を推薦するシステムと考えることができ，基本的な考え方は協調フィルタリングを参考としている．

また松尾らは Web から研究者個人間の繋がりとそれに付随するさまざまな情報を Web から抽出し，研究者の効果的な共同研究の促進のための Polyphonet を作成した [11][12]．この研究では組織内のユーザ，研究者の繋がりを抽出することで新たな連携を支援するものであるが，連携したときにどのような研究テーマが実行可能かを示すまでには至っていない．

3　共著予測の手法

論文ネットワークの共著予測はグラフにおけるリンクマイニングで扱う問題の 1 つであり，著者をノード，共著をリンクとした共著ネットワークから，ある時点のネットワークを元に，今後のリンクを予測する問題として定式化される [9][10]．ここでは協調フィルタリングを用いて，共著予測を行なう方法について述べる．ネットワークの情報から共著を予測する方法では下記の 2 つの方法が考えられる．

- 似た著者同士に新たに共著リンクが発生されやすい
- 同じ分野の著者が選んだ共著者と新たに共著が発生されやすい

本稿では後者の方法を協調フィルタリングにより実装した．前者の方法は，著者間の類似度を求め，類似度の高いペアほど将来新たに共著しやすいと考える方法である．Nowell らはこの著者間の類似度として，Common neighbor, Adamic/Adar, Preferential attachment, Distance の指標を用いて共著リンクの予測を行なった．これら指標については次節の実験方法で述べる．

協調フィルタリングを用いた共著の予測手法は下記の手順で行なう．

(1) データ収集
論文データベースより，年度ごとの論文情報を収集する．

(2) 研究者 A と類似する分野の n 人の研究者グループ A_0 を求める．
研究者 A, B 間の類似度 $\mathrm{sim}(A,B)$ の計算は，研究者が書いた論文に出現するキーワード間の共起頻度の強さを用いた．ここでは，研究者 A, B の書いた論文に出現する単語の AND, OR をとったときの値をそれぞれ $|a \cap b|$，$|a \cup b|$ とする．

$$\mathrm{sim}(A,B) = \frac{|a \cap b|}{|a \cup b|} \tag{1}$$

(3) 類似研究者グループ A_0 の共著者の中で研究者 A が共著していない研究者の共著予測値を求める．
式 (2) は A の類似研究者 $a0 \in A_0$ と共著しており，かつ A と共著していない研究者 b_k すべてに対して，予測値 R_{AB_i} が計算される．

$$R_{AB_i} = \overline{r_i} + \frac{\sum_{a0 \in A_0} \mathrm{sim}(A,a0) \cdot (r_{a0}(b_k) - \overline{r_0})}{\sum_{a0 \in A_0} |\mathrm{sim}(A,a0)|} \tag{2}$$

ただし $\overline{r_0}$ は b_k の共著回数の平均である．共著予測式は (1) 式の類似度で他人との共著回

数に重み付けを行なうことにより，ずれの度合を算出するものである．したがって，予測値は類似度が高い研究者には大きく影響され，類似度の低い研究者にはほとんど影響されない．

(4) 共著予測候補者 B_i を共著予測値が高い順に出力する．
最終的に共著予測値 R_{AB_i} の上位 m 人を選ぶ．m は実際に新規に共著した人数から求める．

4　実験

提案手法の精度を確かめるため，CiNii 提供の情報処理学会の発行する論文データを用いて実験を行なった．

4.1　実験データ

提案手法の共著予測精度を確かめるために CiNii から情報処理学会が 2003 - 2007 年の間に発行した論文データを用いた．表 1 に実験に使ったデータを示す．

表 1: 実験データ 共著ネットワーク

年	著者数	共著数	新規共著数
2003	5337	26664	-
2004	5379	23828	7758
2005	5391	25592	8398
2006	5742	29099	9649
2007	5825	30629	10111

4.2　実験方法

共著予測の実験は下記の手順で行なった．

1. 学習期間を各年に発行された論文から得られた共著ネットワークとし，学習期間のネットワークを N_t，共著予測評価期間のネットワークを $N_{t'}$ を作成する $(t < t')$.
2. N_t の全ての研究者について，共著予測の手法を適用し，共著予測値 R_{AB_i} を求める．
3. 共著予測値の高い研究者と新たに共著する可能性が高いとし，スコアの高い研究者 m 人を選ぶ
4. $N_{t'}$ と比較し，選んだ研究者が実際に共著されていれば予測が正しいので，下記の式から精度を求める．

$$\text{Accuracy} = \frac{m'}{m}$$

m' は一致した共著者数

ここで比較のために Nowell らが用いた手法を使用した．Nowell らは類似度が高い研究者間に新たな共著が発生すると考え，下記の Common neighbors, Jaccard, Adamic/Adar, Preferential attachment の 3 つの類似度算出方法を用いた．この類似度が高い研究者 m 人を共著予測研究者として，提案手法と同様に精度を求めた．

1. Common neighbors

$$\text{sim}(A, B) = |\Gamma(A) \cap \Gamma(B)|$$

2. Jaccard

$$\text{sim}(A, B) = \frac{|\Gamma(A) \cap \Gamma(B)|}{|\Gamma(A) \cup \Gamma(B)|}$$

3. Adamic/Adar

$$\text{sim}(A, B) = \sum_{C \in \Gamma(A) \cap \Gamma(B)} \frac{1}{\log|\Gamma(C)|}$$

4. Preferential attachment

$$\text{sim}(A, B) = |\Gamma(A)| \cap |\Gamma(B)|$$

ここで $\Gamma(A)$ は研究者 A の共著集合を表わす．

それぞれ Common neighbors は共通の共著者を多く持っているほど，高い値を示すものであり，Jaccard は正規化された Common neighbors であり，共通の共著者の割合が高いほど高い値を示す．同様に Adamic/Adar は小数の共著者をもつ研究者を共通の共著者に持つほど，Preferentialattachment は共著者が多い研究者ほど，高い値を示す．

4.3　実験結果

共著予測の結果を図 3 に示す．ベースラインとしてランダムに共著者を予測した場合を rand とした．

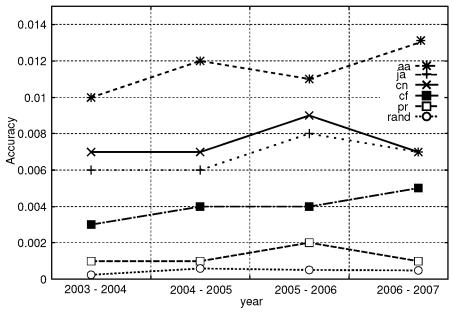


図 3: 共著予測精度

図 3 より，どの手法も rand の 2-10 倍程度の精度を示していることがわかる．共著予測精度の高い順に，Adamic/Adar, Common neighbors, Jaccard, 提案手法 (CF)，Preferential attachment となっており，提案手法は共著予測について特に高い値を示すことはなかった．Nowell らの実験においても Adamic/Adar が最も良い精度を示していることから Adamic/Adar はリンク指標という点において比較的高精度な予測が可能であることがわかる．

4.4　考察

考察として，提案手法で共著予測を行なったときに共著予測された研究者がこれまでにやった発表した論文などを見れるビューアを作成した (図 4)．

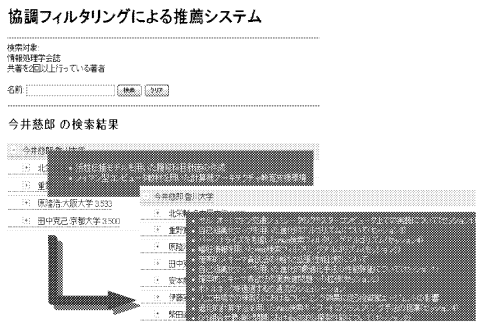


図 4: 共著予測した研究者の論文一覧

このビューアからは研究者が書いた論文データが見ることができる．このデータは実験で利用した論文データだけであるため，予測された共著者の全ての論文を見ているわけではないが，予測された研究者が出した論文一覧を見ることでどの分野と親和性が高いのかということを知ることができる．この共著予測された研究者のリストは，類似する研究者が良く共著する研究者であり，研究者の共同研究動向が反映されたものである．

5　今後の課題

本研究は研究者の行なっている研究内容や研究者同士の共著関係を用いて，共著予測を行なう方法について述べた．共著予測精度の点において，先行研究に比べ低いものの，研究者にとって類似分野の研究者の共同研究の振舞いを元に得た共著の予測は，単に共著するひとを知るだけでなく今後どのような分野の人と共同研究すれば良いのか，どういった分野が親和性が高いのかといったことを知ることができると考えられる．今後，本手法を研究者にとって効果的な共同研究支援のためのツールとして開発していきたいと考えている．

今後の課題として，共著予測を人を単位にするのではなく，親和性の高い共同研究可能な分野といって示すための方法を示すこと，精度以外の評価手法の検討すること，論文データベースにおける名寄せの処理などがあげられる．

参考文献

[1] 神田由美子，他: 科学技術指標，文部科学省科学技術政策研究所, 2008.7.

[2] 孫 媛, 西澤 正己, 根岸 正光: 日本の論文の共著関係からみた産学連携の現状分析，情報知識学会誌, Vol.15, No.2,pp. 89-92, 2005.

[3] 芳鐘 冬樹, 影浦 峡: 共著ネットワークの分析 視点と手法, 図書館情報学会, 第 49 回日本図書館情報学会研究大会, p. 55-58, 2001.

[4] Nowell, D. L., Kleinberg, J.: The Link Prediction Problem for Social Networks, Proc of 12th International Conference on Information and Knowledge Management, pp.556-559, 2003.

[5] 長尾 純子, 佐藤 弘行, 菊池 俊一: 研究開発支援総合ディレクトリ (ReaD). 情報管理. Vol. 45, No. 12, (2003), 891-894.

[6] 廣川佐千男, 関隆宏, 安元裕司, 山田泰寛: 教員データに対する多面的検索システム，情報処理学会研究報告 DBS Vol.2005, No.11, pp.665-672, 2005.

[7] 関 隆宏，安元 裕司，廣川 佐千男: 教員データにおける高頻度語 (語の統計量とその応用)，情報処理学会自然言語処理研究会報告， Vol.2005, No.22, pp.1-8, 2005.

[8] 田中省作, 関隆宏, 石野明, 金丸玲子, 杉本典子, 竹田正幸, 廣川佐千男: 大学経営における大学評価システムの活用, 第 67 回情報処理学会全国大会，2005.

[9] 鹿島 久嗣, 安倍 直樹: ネットワーク構造の確率的な時変モデルに基づく教師ありリンク予測, 人工知能学会論文誌, Vol. 22, No. 2, pp.209-217, 2007.

[10] 鹿島 久嗣: ネットワーク構造予測, 人工知能学会論文誌, Vol.22, No.3, pp.344-351 2007.

[11] 松尾豊, 石黒周，松原仁, 橋田浩一，中島秀之: 効果的な共同研究を支援するための Web からの研究者ネットワーク抽出，情報処理学会，情報学シンポジウム, 2005

[12] 松尾 豊, 石黒周, 松原 仁, 橋田浩一, 中島秀之: 効果的な共同研究を支援するための Web からの研究者ネットワーク抽出，情報学シンポジウム，2005

[13] 角田雅照，大杉直樹，門田暁人，松本健一，佐藤慎一，”協調フィルタリングを用いたソフトウェア開発工数予測方法,” 情報処理学会論文誌, Vol.46, No.5, pp.1155-1164, 2005.

[14] 大杉直樹， 門田暁人， 松本健一， 森崎修司: 協調フィルタリングに基づくソフトウェア機能推薦システム, 情報処理学会論文誌，Vol.45, No.1, pp.267-278, 2004.

[15] 市井誠， 山本哲男， 横森励士， 井上克郎: ソフトウェア部品推薦のための協調フィルタリング手法の提案と実現, 電子情報通信学会技術研究報告, SS2004-15, 2004

[16] 土方嘉徳: 情報推薦・情報フィルタリングのためのユーザプロファイリング技術，人工知能学会誌，Vol.19, No.3, pp.365-372, 2004.

[17] 土方嘉徳: 嗜好抽出・情報推薦の基礎理論：1）嗜好抽出と情報推薦技術，情報処理学会誌, Vol.48, No.9, pp.957–965, 2007.

[18] Herlocker, J., et al.: Evaluating Collaborative Filtering Recommender Systems, *ACM Transactions on Information Systems*, Vol.22, No.1, pp.5-53, 2004.

世代間共同研究による知識移転のモデル化

Modeling Knowledge Transfer through Inter-Generational Research Collaboration

福田　毅哉[1*]，梅本　勝博[2]
Takaki FUKUDA, Katsuhiro UMEMOTO

*1 東京大学物性研究所/北陸先端科学技術大学院大学知識科学研究科
The Institute for Solid State Physics, The University of Tokyo / School of Knowledge Science, Japan Advanced Institute of Science and Technology.
〒277-8581 千葉県柏市柏の葉5-1-5
E-mail: fukuda@issp.u-tokyo.ac.jp
2 北陸先端科学技術大学院大学知識科学研究科
School of Knowledge Science, Japan Advanced Institute of Science and Technology.
〒923-1292 石川県能美市旭台1-1
E-mail: ume@jaist.ac.jp

科学知はどのようにして移転されているのか？　共同研究では知識が共有されること、また、共同研究が世代交代することにより、直接接していない研究者にも知識が移転される可能性があることを示す。

How is scientific knowledge transferred? To create new knowledge, research collaboration plays a role not only by sharing knowledge but also transferring knowledge indirectly to other scientists.

キーワード: 共同研究, 知識移転, 暗黙知
Research Collaboration, Knowledge Transfer, Tacit Knowledge

1 はじめに

21世紀は知識社会といわれる。今までより知識が社会にとって大きな影響を持つ社会であり、知識についての研究が必要である。

1.1 知識に関する先行研究

知識創造についての研究は、経営学の分野でナレッジ・マネジメントとして盛んに行われている。個人と集団の間で形式知と暗黙知の変換により知識を創造するSECIモデル[1]が提唱された。

その後、知識共有と創造に関して、知識を注意深くコード化してデータベースなどを通じて再利用する「コード化戦略」と、個人に属する暗黙知を個人間のコミュニケーションをもとに進める「個人化戦略」が区分された[2]。また、知識創造と文脈を共有する場とリーダーシップ[3]などの研究が行われてきた。

1.2 知識移転に関する先行研究

知識移転は、組織にとって重要なテーマである。

業務と文脈の類似性や業務の質から、知識移転は、連続移転、近接移転、遠隔移転、戦略的移転、専門知移転と区分されている[4]。

知識移転の方法は、組織の文化に適合していなければならず[5]、多国籍企業内部で国際間の知識移転をする際には、異文化マネジメントが欠かせない[6]ということが言われている。

また、知識移転のコストは、受け手と送り手の間における伝達・受信能力や事前知識のギャップによって決まることがわかっている[7]。

1.3 科学研究に関する先行研究

知識について理解する上で。知を生み出すことを目的としている科学研究についての研究は重要である。

人類学では研究コミュニティと知識成長に関して研究されている[8]。また、認知科学では、科学的発見について、仮説空間と実験空間という2つの空間の探索として捉えられている[9]。

科学研究の知識を産業にどう移転するのかという視点では、共鳴しながらプロジェクトを推進できる場が必要［10]とされている。

本稿では、科学についてのデータ・情報・知識・知恵を含めて科学知とする。

2 モデル提案

2.1 知識創造の状態遷移モデル

知識創造に外部がどのように影響しているのかを考慮するため、個人の行為により、個人の状態が変化するというモデルを示す。

知識創造も状態変化として表す（図1）。

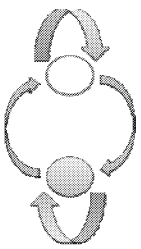


図 1 知識創造の状態遷移モデル

上の○は知識を創造する個人を、下の●は外部環境を表す。矢印は、行為を表す。

2.1.1 個人における知識創造の状態遷移モデル

個人Aがある課題について考え、アイデアを思いついた。その結果起こった状態変化をA→A′とする。

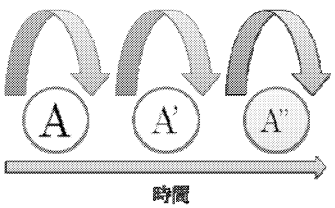


図 2 個人における知識創造の状態遷移モデル

図2では、A′ の状態から、ある行為（アイデアの検証など）を経て知識を生み出した(A″)ことを表している（図2）。

2.1.2 外部とのコミュニケーションについての知識創造の状態遷移モデル

2.1のモデルを拡張し、外部とのコミュニケーションによっても状態変化が起こることを表している。

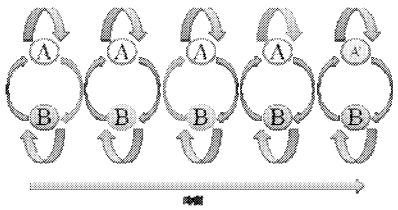


図 3 外界との交わりによって行われる知識創造モデル

図3では、Aが外部Bに働きかけ、Bからの返信を認識することにより、知識創造した(A′)ことを示している（図3）。

2.2共同研究の世代交代と知識移転

2.2.1 共同研究における知

共同研究とは、複数人でテーマを共有し、協力して研究することである。複数の人間の頭脳が集まったときには個人では起こらないようなアイデアの相乗効果が生まれることがある[4]。共同研究の結果、科学知の発展が認められたときには共著論文として出版される。

実験データなどの形式知や、実験のコツなどの暗黙知も人と人とが接することにより共有され、新しい知識の創造が行われている。

2.2.2 共同研究の世代交代

研究者SがT,Uと共同研究Rを行った。その後、SはV,X,Wと共同研究R′ を行った。このように共同研究の相手が替わるとき、Sにとって共同研究が世代交代したとする（図4）。

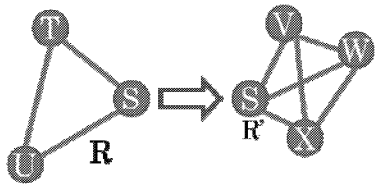


図 4 共同研究の世代交代
●は研究者を表す。リンクは共同研究を行っている関係を表す。

2.2.3 世代間共同研究の知識移転モデル

Sが行った2つの共同研究RとR′ には、共同研究者の重なりは見られない。しかし既に行った共同研究RにおいてTやUの知をSが共有した場合、R′ でSを介してTやUの知をV,W,Xに移転する可能性がある（図5）。

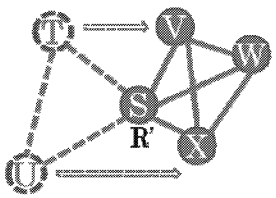


図 5 世代間共同研究による知識移転
点線は過去の共同研究を、実践は現在の共同研究を表す。

3 まとめ

現在の多くの科学知は共同研究により生みだされている。そこでは研究者同士の接触により知が共有される。本稿では、共同研究をする相手を替えることにより、直接接していない研究者にも知が移転される可能性のあることを示した。

本稿はモデルの提案であり、検証が必要である。共同研究の成果として示される共著論文の共著者ネットワークの分析を行うことにより、知識移転の構造を理解できる。また、サーベイやアンケートにより個々の共同研究の推移を分析することで、実際にどのような知識が移転しているのか検証することができる。

謝辞

本稿をまとめるにあたり東京大学物性研究所技術専門職員の矢田裕行さんにご議論いただいたことを記し感謝を表する。

参考文献

[1] 野中郁次郎；竹内弘高（梅本勝博訳）:「知識創造企業」,東洋経済新聞社, 1996.

[2]Hansen, M. T.: Nohria, Nitin : Tierney, Thomas : “What’s your strategy for managing Knowledge? ,” Harvard Business Review, Mar.-Apr. ,pp.106-116, 1999

[3] 遠山亮子；野中郁次郎：「『よい場』と革新的リーダーシップ：組織的知識創造についての試論」, 一橋ビジネスレビュー , Vol.48, Sum.-Aut., pp.1-13, 2000.

[4] Dixon, Nancy M.（梅本勝博：遠藤温：末永聡訳）:「ナレッジ・マネジメント 5つの方法 課題解決のための『知』の共有」, 生産性出版, 2003.

[5]Davenport, Thomas H. : Prusak, Laurence（梅本勝博訳）:「ワーキング・ナレッジ 『知』を活かす経営」, 生産性出版, 2000.

[6]高橋俊一；「多国籍企業内部の国際間知識移転における現地利害関係者管理の戦略的重要性に関する研究：異文化マネジメントの視点から」,立教経済学研究, Vol.61, No.3, pp.145-170, 2008

[7] 犬塚篤：「組織内ネットワークの構築と知識共有(<特集>知識継承）」,人工知能学会誌, 22(4),pp.472-479, 2007.

[8] Crane, Diana（津田良成監訳, 岡沢和世訳）:「見えざる大学」,敬文社 ,1979

[9] 齋藤 ひとみ;「Web情報探索における認知プロセスの理解とその応用」, 情報知識学会誌, Vol. 16, No. 4,pp.51-62, 2006.

[10] 八代英美；小林俊哉：「ロシアの基礎研究から日本産業への知識移転」, 研究技術計画, Vol.21(3/4), pp.243-251 ,2007.

完全合理性と限定合理性

Complete rationality and definitive rationality

安平哲太郎*

Tetsutaro YASUHIRA

***産業技術総合研究所　Advanced　Institution of Science and Technology**

〒305-8561　茨城県つくば市東 1-1-1　中央第 2　E-mail：t-yasuhira@aist.go.jp

概要

情報関連装置への入力以前の問題として、客観的世界に働いている完全合理性とそれを認識する人間の限定合理性との関係を議論する。完全合理性は人間が認識するという立場からは、「2 つの対象を人間がある観点から見た時に、本来同じと認識すべき事、異なると認識すべき事」と表現できる。これは、同時点の別の 2 つの対象や、ある時点での対象と経時変化を経た後の同じ対象との間で成立し、異なり方は部分と全体の間で、さらに、経時変化を経る間に外部から働く作用と部分あるいは全体との間で一対一対応である。一方、人間の認識は不完全であり、それ故に限定合理性といわれる。このうち、特に異なると認識できていない時、対立や矛盾が生ずる事をパラドックスを用いて明らかにする。そして、この限定合理性を完全合理性に近づけてゆくためには、部分と全体との関係について問題意識を持つ事が重要で、これが概念分析の本質であることを示す。

As one of problems before input into information- related-devices, there is the relation between a complete rationality working at the objective world and a definitive rationality of human recognizing it. The rationality working objectively can be expressed as what human had better originally recognize as a same or as a different when he observes two objects if we express it from a point of view he recognizes. This is true between two other objects at the same time or between one object at one time and the same object after some time passed and the difference is one to one between a part and a whole at the same time and before and after some time passes it is one to one between the operation working from outer and a part or a whole. On the other hand, because recognition of human is incomplete, he haven`t been able to recognize as a same or a different. This is called as the definitive rationality. We make it clear that he feels opposition and contradiction when he cann`t recognize as a different what he should recognize originally as a different. In order to approach this definitive rationality to the complete rationality, it is important to have the awareness of the issues and this is the essence of the concept analysis.

キーワード：完全合理性、限定合理性、矛盾、パラドックスの解消、概念分析

Complete rationality, Definitive rationality, Contradiction, Resolution of paradox, Concept analysis

1 はじめに

根岸正光会長[1)]が米田幸夫初代会長の言葉「入力以前の問題」を引用して、コンテンツ（内容）が情報知識学においては本質的に重要であり、古くて新しい問題であることを指摘している。そこで、ここでは、入力以前の、そして、情報知識の内容に関する問題として合理性を取り上げることにする。

合理主義という言葉を万有国民百科事典で調べてみると、「非合理、偶然的なものを排し、理性的、論理的、必然的なものを尊重する立場」とある。そこで、次に理性を調べてみると、「物事を正しく判断する力。また、真と偽、善と悪を識別する能力。ときには美と醜を識別する働きまで理性に帰せられることがある。」とある。したがって、まず、理性は基本的には、真という視点から見て真に属するか属さないか、善という視点から見て善に属するか属さないかという具合に、客観的世界を認識においてある視点から見て同じか異なるかという区別をする事だと考えられる。したがって、理性的というのは客観的世界を本来認識されるべき状態として好悪の感情を離れて冷静に異種混合を排して認識する事だと考えられるし、論理的というのは異種混合を排して表現する事だと考えられる。そしてまた、必然的というのは一つの原因に対して一つの結果が対応することであると考えられる。そうすると、合理主義というのは「客観的世界の認識や表現において異種混合を排し、一対一対応の因果関係を尊重する立場」と考える事が出来る。一方、表現もまた、表現者の認識の投影である事を考えると、結局、認識において異種混合を排するという観点から合理性を考える事が出来るのではないかと思われる。そこで、以下、この観点から合理性を検討してゆく事にする。

2 合理性とは

合理性とは、上記の検討から基本的には人間の認識における異種混合を排除することを尊重する立場だと考えられる。しかし、認識がまずは客観的世界をあるがままに区別して受け止める事だとするなら、我々の認識を離れた客観的世界の構造がそうなっているという仮説が存在している事になる。そこで、その仮説を人間の認識という観点から表現してみる事にする。

この時、客観的世界に同時に存在する2つの個別の対象と、1つの対象と、それが経時変化を経た後の対象との2つの場合が考えられる。

Ⅰ．同時に存在する2つの個別の対象に対して、

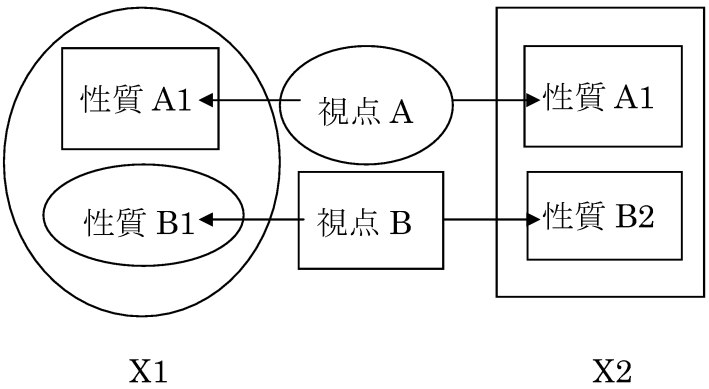
(1) 2つの対象がどの視点から見ても、本来全く同一の性質および関係を持つと認識すべき時、その2つの対象は同一である。

(2) 2つの対象がある視点から見て、

本来共通する性質および関係を持つ部分と、別の視点から見て異なる性質および関係を持つ部分とがあると認識すべき時、その２つの対象は異なっており、したがって、必然的に部分の異なり方と全体の異なり方は１対１対応である（図１）。

この「認識すべき状態」が客観的に働いている合理性であり、これは我々人間の認識の不完全さには依存しないので完全合理性という事にする。

図１．ある２つの対象がある視点から見て同一であると認識すべき性質の部分と別の視点から見て異なると認識すべき性質の部分とに分けられる時、その２つの対象は異なっていると認識しなければならない。

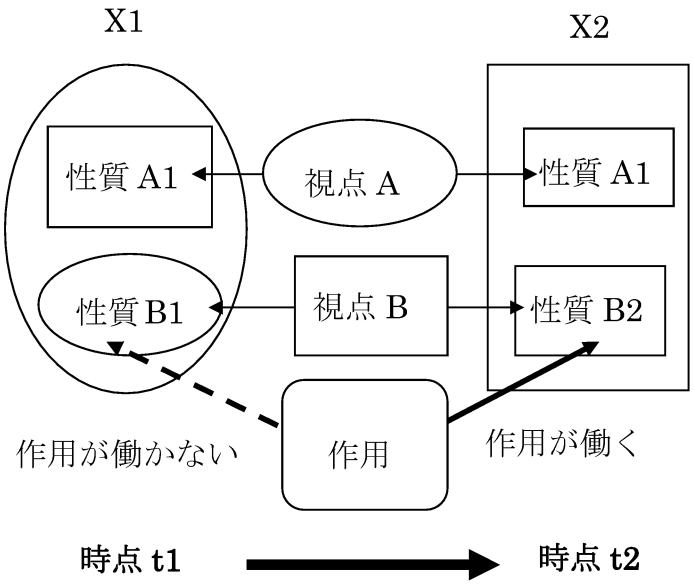


Ⅱ．１つの対象の経時変化による状態変化に対して、

(1) １つの対象が時間経過を経ても、どの視点から見ても全く同一の性質および関係をもつと認識すべき時、外部から何らの作用がなかったと考え、同じ対象と考える。

(2) １つの対象が時間経過を経た時、初期状態とは異なる対象となったと認識すべき時、その間に外部から何らかの作用があったと考える。したがって、必然的に外部からの作用の異なり方と作用が働いた部分および全体の異なり方とは１対１対応である。（図２）

図２．１つの対象が、経時変化を経て、異なると認識すべき対象となるなら、ある視点から見て同一と認識すべき性質の部分と、その間に外部から作用を受けて、別の視点から見て異なると認識すべき性質の部分とがある。



以上、説明した合理性は、我々人間の主観を離れた客観的世界で働いている完全な合理性を、我々人間が認識するという立場から記述したものである。それに対して、それを認識する我々人間の認識は不完全であるから、同じ時点の２つの対象、あるいは、異なる時点の一つの対象の視点Ａから見た部分的な性質の類似性Ａ１や視点Ｂから見た部分的な性質の相違性Ｂ１、Ｂ２を認識できていない事がある。そのような我々の認識の不完全性による合理性を限定合理性といっている。ここでは、２つの命題に対してある観点から同じであるべきものを同じとみていて、かつ、別の観点から本来異なるべきものを区別出来ずに同じとみている時、そして、その観点さえ意識していない時、どういう事が起きるか2つのパラドックスを用いて検討する。

3 矛盾の本質とその解決

ここでは、我々が矛盾を感じる時の表現と認識の構造を説明し、その表現では記述されていない視点を導入する事によって矛盾を解決できる事を明らかにする。そこで、まず、パラドックスと言われている表現を分析して共通の性質を見出し（これは客観的に認める事が出来るものである。）、つぎに、この共通の性質をすべて満足するように、すなわち、騙されるところは騙され、認識すべきところは認識したらどのように感じるかを考えていく事にする。具体例としてゼノンの「英雄アキレスと亀の競走」[2)]と韓非子の故事「矛と楯」というパラドックスを取り上げる。

本題に入る前に、本論文を理解するために必要な程度の言葉の定義を簡単にしておきたい。まず、表現とは、表現の読み手がある意味を認識する事を表現者の意図としてあらわした文章という事にする。また、意味とは、言語表現された「ある対象に関連して考えた方が良い事」、「ある対象に関連して考えるべき事」と定義する。そうすれば、意味を、客観的に存在している対象を考えた時、その対象をある観点から見て価値判断した性質と考える事も出来る。したがって、様々な観点から対象を見て価値判断した性質の集合がある対象に対するコンテンツということになる。すなわち、コンテンツはその対象に付与された様々な意味の集合である。そして、客観的に存在する対象として表現を考えた時、表現者の意図という観点から判断した意味を考えることができる。したがって、表現Ａに対して表現者の意図通りの意味「Ａ」を意図通りの意味という事にする。そして、この意味が現実にあり得る時、現実的意味といい、現実にあり得ないとき、空想的意味という事にする。そして、この意味を構成する単位を概念という事にする。また、1 つの命題は、1 つの完結された表現であるが、それを構成する部分を部分表現という事にする。

3．1 人間に矛盾を感じさせる表現の構造

(1) ある視点から見た具体的性質を無視して記述されている部分表現Ａ

(2) もし、部分表現Ｂが現実的意味を持つなら、その視点に関して具体的性質が表現Ａとは異ならなければならない事を無視して記述されている部分表現Ｂ

(3) 部分表現Ａと部分表現Ｂにおける異なる具体的性質を記述しない事によって、部分表現Ａと同一の性質であるかの様に錯覚させ現実にあり得ない意味にしておきながら、なおかつ部分表現Ａの意図通りの意味と部分表現Ｂの意図通りの意味とが現実に両立していると認識せざるを得ない構造

例）部分表現Ａと部分表現Ｂとが前提－結論、原因－結果、一様な連続な図形などとして表現される。

以下に上の各性質について、具体例で見てゆく。

具体例　英雄アキレスと亀の競走

アキレスと亀とが競走する。アキレスは有名な俊足の持ち主であるから、ハンディキャップをつけて、亀より後の位置から同時に出発する。さて、アキレスが亀を追い越すためには、まず、亀の出発点に立たなければならない。しかし、その時は、亀はもっと先の地点にいる。そして、アキレスがその地点に到達したときには、亀はさらに先の地点にいる。アキレスが亀の地点に到達するたびに、いつでも亀は少し先に進んでいる。だから、アキレスは、いつまでたっても亀に追いつけない。アキレスがいくら早く走っても無駄であるし、亀はいくら遅くとも、休まずに進み続ける限り、決して追いつかれることはない。

(0)視点となっている現実の性質

アキレスの歩幅

(1)具体的性質を無視して記述されている部分表現Ａ、(以下[　]内は無視している具体的性質)

アキレスが亀を追い越すためには、まず、[一定の歩幅で]亀の出発点に立たなければならない。しかし、その時は、亀はもっと先の地点にいる。そして、アキレスがその地点に到達したときには、亀はさらに先の地点にいる。

(2)もし、部分表現Ｂが現実的意味を持つなら、その視点に関して具体的性質が部分表現Ａとは異なる事を無視して記述されている部分表現Ｂ(以下[　]内は無視している具体的性質)

[アキレスと亀の距離がアキレスの歩幅に等しくなってからアキレスの歩幅をだんだん狭めて]アキレスが亀の地点に到達するたびに、いつでも亀は少し先に進んでいる。

(3)部分表現Ａと部分表現Ｂにおける異なる具体的性質を記述しない事によって、部分表現Ａと同一の性質であるかのように錯覚させ現実にあり得ない意味にしておきながら、なおかつ部分表現Ａの意図通りの意味と部分表現Ｂの意図通りの意味とが現実に両立していると認識せざるを得ない構造

部分表現Ａと部分表現Ｂとがそれぞれ現実的意味を持つためにはアキレスと亀の距離がアキレスの歩幅に等しくなってからアキレスの歩幅が狭まることが必要であるにもかかわらずそれを記述せず、いつでも同じ歩幅でアキレスが亀の地点に到達する事が可能であるかのように記述している。また、「競走する」とか「俊足の持ち主」、「するたびに、いつでも」「亀に追いつけない」、「いくら早く走っても」という表現を用いたりして歩幅が変わっていないという印象、さらに「Ａ」と継続した時間上にあり、従って原因に対する結果という印象を強めている。そして、「亀はいくら遅くとも、休まずに進み続ける限り」という表現を用いることによってアキレスの歩幅に関心が行かないようにしている。

具体例　矛と楯

どの楯をも貫き通す矛とどの矛をも通させない楯がある。この矛と楯とが戦ったらどうなるか。

(0)視点となっている現実の性質

矛と楯が存在する時間範囲

(1) 具体的性質を無視して記述されている部分表現Ａ（以下[　]内は無視している具体的性質）

[ある時間範囲で]どの楯をも貫き通す矛がある。

(2) もし、部分表現Ｂが現実的意味を持つなら、その視点に関して具体的性質が部分表現Ａとは異なることを無視して記述されている部分表現Ｂ（以下[　]内は無視している具体的性質）

[それとは異なる時間範囲で]どの矛をも通させない楯がある。

(3)部分表現Ａと部分表現Ｂにおける異なる具体的性質を記述しない事によって、部分表現Ａと同一の性質であるかのように錯覚させ現実にあり得ない意味にしておきながら、なおかつ部分表現Ａの意図通りの意味と部分表現Ｂの意図通りの意味とが現実に両立していると認識せざるを得ない構造

現実にあり得るなら時間範囲が異ならなければならない事を記述せず、且つ、「と」と言う並置を用いたり、「とが戦ったら」という言葉を用いたりして、同時には存在し得ない事象が同時に存在し得ているかのように錯覚させている。

３．２ 人間に矛盾を感じさせる認識の構造

(1) ３．１「人間に矛盾を感じさせる表現の構造」の(3)によって指定されている関係の為に、部分表現Ａと部分表現Ｂとで具体的性質が変わらなければ現実にはあり得ないにもかかわらず、具体的性質が変わらないと錯覚したまま現実にあり得るような感じがすると共に、

(2) ３．１「人間に矛盾を感じさせる表現の構造」の(3)によって指定されている関係に対して、経験より部分表現Ａより実際に具体的性質が変わらない時に現実に成立し得る意味 not「Ｂ」が潜在意識下で誘起され、現実であると錯覚している意図通りの空想的意味「Ｂ」と経験による現実的意味 not「Ｂ」とが重なって違和感を感ずる。

(3)意味「Ｂ」と意味 not「Ｂ」とも(1)、(2)の関係より現実でなければならないと思っているにもかかわらず、現実には両立し得ない事も認識しているから、違和感を解消しなければならないと思う。

(4)そこで今度は、その違和感を解消しようとして部分表現Ｂに対して部分表現Ａが現実に成立するかどうか見ようとする、

(5)「人間に矛盾を感じさせる文章の構造」の(3)によって指定されている関係に対して、経験より部分表現Ｂより実際に具体的性質が変わらない時に現実に成立し得る意味

not「A」が潜在意識下で誘起され現実であると錯覚している意図通りの空想的意味「A」と経験による現実的意味 not「A」とが重なって違和感を感ずる。

(6) 意味「A」と意味 not「A」とも(4)、(5)の関係より現実でなければならないと思っているにもかかわらず、現実には両立し得ない事も認識しているから、違和感を解消しなければならないと思う。

以下 (1) から繰り返す。

パラドックスによっては、(3)で(1)へ戻るパラドックスもある。このアキレスと亀の競走の場合がそうである。

以下に、上の各性質について具体例でみてゆく。

具体例　英雄アキレスと亀の競走

(1) ３．１「人間に矛盾を感じさせる表現の構造」の(3)によって指定されている関係の為に、部分表現Ａと部分表現Ｂとがアキレスの歩幅という視点に関して具体的性質が変わらなければ現実にはあり得ないにもかかわらず、具体的性質が変わらないと錯覚したまま現実にあり得るような感じがすると共に、

(以下[　]内は意識していない)

部分表現Ａ：

アキレスと亀とが競走する。アキレスは有名な俊足の持ち主であるから、ハンディキャップをつけて、亀より後の位置から同時に出発する。さて、アキレスが亀を追い越すためには、まず、[一定の歩幅で]亀の出発点に立たなければならない。しかし、その時は、亀はもっと先の地点にいる。そして、アキレスがその地点に到達したときには、亀はさらに先の地点にいる。

部分表現Ｂ：

[アキレスと亀の距離がアキレスの歩幅に等しくなっても一定の歩幅で]アキレスが亀の地点に到達するたびに、いつでも亀は少し先に進んでいる。だから、アキレスは、いつまでたっても亀に追いつけない。

(2) ３．１「人間に矛盾を感じさせる表現の構造」の(3)によって指定されている関係に対して、経験より部分表現Ａより実際に具体的性質が変わらない時に現実に成立し得る意味 not「B」が潜在意識下で誘起され現実であると錯覚している意図通りの空想的意味「B」と経験による現実的意味 not「B」とが重なって違和感を感ずる。**(以下[　]内は意識していない)**

文章から現実にあり得ると錯覚させられている空想的意味：

亀より後の位置から同時に出発する。さて、アキレスが亀を追い越すためには、まず、[一定の歩幅で]亀の出発点に立たなければならない。しかし、その時は、亀はもっと先の地点にいる。そして、アキレスがその地点に到達したときには、亀はさらに先の地点にいる。**(意図通りの現実的意味「A」)**

[アキレスと亀の距離がアキレスの歩幅に等しくなっても一定の歩幅で]アキレスが亀の地点に到達するたびに、いつでも亀は少し先に進んでいる。だから、アキレスは、いつまでたっても亀に追いつけない。**(意図通りの空想的意味「B」)**

経験による現実的意味：
亀より後の位置から同時に出発する。さてアキレスが亀を追い越すためには、まず、[一定の歩幅で]亀の出発点に立たなければならない。しかし、その時は、亀はもっと先の地点にいる。そして、アキレスがその地点に到達したときには亀はさらに先の地点にいる。**(意図通りの現実的意味「A」)**
―>それでもいずれは[一定の歩幅で]アキレスは亀を追い越すはず。
(経験による意図から離れた現実的意味 not「B」)

(3) 意味「B」と意味 not「B」とも(1)、(2)の関係より現実でなければならないと思っているにもかかわらず、現実には両立し得ない事も認識しているから、違和感を解消しなければならないと思う。

具体例　矛と楯

(1) 3．1「人間に矛盾を感じさせる文章の一般的構造」の(3)によって指定されている関係の為に、矛と楯が存在する時間範囲という視点の存在とその視点が変化していることを気づかせずに、部分表現Aと共に部分表現Bが現実に成立するような感じがすると共に、(以下[　]内は意識していない)

部分表現A：
【ある時間範囲に】どの楯をも貫き通す矛
部分表現B：
と【その同じ時間範囲に】どの矛をも通させない楯がある。

(2)「人間に矛盾を感じさせる文章の一般的構造」の(3)によって指定されている関係に対して、経験により意図通りの現実的意味「A」より意図とは離れた現実的意味 not「B」が潜在意識化で誘起され、意図通りの空想的意味「B」と意図から離れた現実的意味 not「B」とが重なって違和感を感ずる。(以下[　]内は意識していない)

文章から現実にありうると錯覚させられている空想的意味：
【ある時間範囲に】どの楯をも貫き通す矛と**(意図通りの現実的意味「A」)**
【その同じ時間範囲に】どの矛をも通させない楯がある**(意図通りの空想的意味「B」)**
経験に基づく現実的意味：
【ある時間範囲に】どの楯をも貫き通す矛があるから**(意図通りの現実的意味「A」)**
【その同じ時間範囲に】どの矛をも通させない楯はない。**(経験による意図**

から離れた現実的意味 not「B」）

(3) 意味「B」と意味 not「B」とも(1)、(2)の関係より現実でなければならないと思っているにもかかわらず、現実には両立し得ない事も認識しているから、違和感を解消しなければならないと思う。

(4)そこで今度は、その違和感を解消しようとして「B」に対して「A」が現実に成立するかどうか見ようとする、

どの矛をも通させない楯（「B」）とどの楯をも貫き通す矛（「A」）とがある。

(5)「人間に矛盾を感じさせる文章の一般的構造」の(3)によって指定されている関係の中で、現実の「B」より現実に成立しうる not「A」が潜在意識化で誘起され、「A」と not「A」とが重なって違和感を感ずる。

文章から現実にありうると錯覚させられている空想的意味：
【ある時間範囲に】どの矛をも通させない楯（「B」）と、【その同じ時間範囲に】どの楯をも貫き通す矛とがある。（「A」）
経験に基づく現実的意味：
【ある時間範囲に】どの矛をも通させない楯がある（「B」）から、【その同じ時間範囲に】どの楯をも貫き通す矛はない。（not「A」）

(6) 意味「A」と意味 not「A」とも(4)、(5)の関係より現実でなければならないと思っているにもかかわらず、現実には両立し得ない事も認識しているから、違和感を解消しなければならないと思う。
以下（1）から繰り返す

３．３ 矛盾の解消

(1) 同じでなければならないはずと感じている２つの命題、
意図通りの現実的意味「A」－＞意図通りの空想的意味「B」と、
意図通りの現実的意味「A」－＞経験による意図から離れた現実的意味 not「B」とを個別に意識する。

(2) 意図通りの現実的意味「A」－＞意図通りの空想的意味「B」を文章から直接感ずる空想的意味、意図通りの現実的意味「A」－＞経験による意図から離れた現実的意味 not「B」を経験から感ずる現実的意味とみなす。

(3) 現実的意味「A」が共通であるが為に本来同じでなければならないと感じているこれら２つの意味を区別するために、意図通りの現実的意味「A」と意図通りの空想的意味「B」に関してそれぞれが別々に現実に成立するとすればどのような性質を現実から見落としているかを考える。

以下に，上の各性質について、具体例でみてゆく。

具体例　アキレスと亀の競走

(1) 同じでなければならないと感じている２つの命題、意図通りの現実的意味「A」－＞意図通りの空想的意味「B」と意図通りの現実的意味「A」－＞経験による意図から離れた現実的意味 not「B」とを個別に意識する。([　]内は意識していない)

現実にありうると錯覚させられている空想的意味：

亀より後の位置から同時に出発する。さて、アキレスが亀を追い越すためには、まず、[一定の歩幅で] 亀の出発点に立たなければならない。しかし、その時は、亀はもっと先の地点にいる。そして、アキレスがその地点に到達したときには、亀はさらに先の地点にいる。**(意図通りの現実的意味「A」)**

[アキレスと亀の距離がアキレスの歩幅に等しくなっても一定の歩幅で]アキレスが亀の地点に到達するたびに、いつでも亀は少し先に進んでいる。だから、アキレスは、いつまでたっても亀に追いつけない。**(意図通りの空想的意味「B」)**

経験による現実的意味：

亀より後の位置から同時に出発する。さてアキレスが亀を追い越すためには、まず、[一定の歩幅で] 亀の出発点に立たなければならない。しかし、その時は、亀はもっと先の地点にいる。そして、アキレスがその地点に到達したときには、亀はさらに先の地点にいる。**(意図通りの現実的意味「A」)**

―＞いずれはアキレスは [一定の歩幅で] 亀を追い越すはず。**(経験による意図から離れた現実的意味 not「B」)**

(2) 意図通りの現実的意味「A」－＞意図通りの空想的意味「B」を文章から直接感ずる意図通りの空想的意味、現実的意味「A」－＞現実的意味 not「B」を経験から感ずる現実的意味とみなす。(以下[　]内は意識していない)

文章から直接感ずる意図通りの空想的意味とみなす：

亀より後の位置から同時に出発する。さてアキレスが亀を追い越すためには、まず、[一定の歩幅で] 亀の出発点に立たなければならない。しかし、その時は、亀はもっと先の地点にいる。そして、アキレスがその地点に到達したときには、亀はさらに先の地点にいる。**(意図通りの現実的意味「A」)**

[アキレスと亀の距離がアキレスの歩幅に等しくなっても一定の歩幅で]アキレスが亀の地点に到達するたびに、いつでも亀は少し先に進んでいる。だから、アキレスは、いつまでたっても亀に追いつけない。**(意図通りの空想的意味「B」)**

経験による現実的意味：

[一定の歩幅で] 亀より後の位置から同時に出発する。さて、アキレスが亀を追い越すためには、まず、亀の出発点に立たなければならない。しかし、

その時は、亀はもっと先の地点にいる。そして、アキレスがその地点に到達したときには、亀はさらに先の地点にいる。**(意図通りの現実的意味「A」)**
―＞それでも経験からいずれはアキレスは［一定の歩幅で］亀を追い越すはず。**(経験による現実的意味not「B」)**

(3)命題「A」が共通であるが為に本来同じでなければならないと感じているこれら２つの命題を区別するために、意味「A」と「B」に関してそれぞれが現実に成立するとすればどのような具体的性質を現実から見落としているかを考える。
（太字斜体部は意識している）

空想的意味に関して、太字斜体部のように視点を修正すれば、現実に在り得る：
アキレスが亀を追い越すためには、まず、**一定の歩幅で**亀の出発点に立たなければならない。しかし、その時は、亀はもっと先の地点にいる。そして、アキレスがその地点に到達したときには、亀はさらに先の地点にいる。**(意図通りの現実的意味「A」)**
そして、***アキレスと亀の距離がアキレスの歩幅に等しくなるぐらいからアキレスの歩幅をだんだん狭めてゆけば***アキレスが亀の地点に到達するたびに、いつでも亀は少し先に進んでいる。だから、アキレスはいつまでたっても亀に追いつけない。**(意図とは離れた現実的意味「B‘」)**

経験による現実的意味：
アキレスが亀を追い越すためには、まず、**一定の歩幅で**亀の出発点に立たなければならない。しかし、その時は、亀はもっと先の地点にいる。そして、アキレスがその地点に到達したときには、亀はさらに先の地点にいる。**(意図通りの現実的意味「A」)**
そして、**アキレスと亀の距離がアキレスの歩幅に等しくなるぐらいからでもアキレスの歩幅が変わらなければ、**アキレスは亀を追い越す。**(経験による現実的意味not「B」)**

以上の議論から分かるとおりアキレスの歩幅（という視点）に関して、歩幅は変わらないことを強く印象付けて、しかも、その事を意識させないようにしながら、実際には「A」と「B」とで歩幅を変えた議論を進め、しかもそれが、それなりに実現可能なものだから、経験から生ずる歩幅は変わらないとした場合の結論との間で矛盾を感じるのである。

具体例　矛盾
(1)同じでなければならないはずと感じている２つの命題、
意図通りの現実的意味「A」－＞意図通りの空想的意味「B」と
意図通りの現実的意味「A」－＞経験による意図から離れた現実的意味 not「B」とを個別に意識する。
現実にあり得ると錯覚させられている空想的意味：
【ある時間範囲に】どの楯をも貫き通

す矛がある。**(意図通りの現実的意味「A」)**
そして【同じ時間範囲に】どの矛をも通させない楯がある。**(意図通りの空想的意味「B」)**
経験に基づく現実的意味：
【ある時間範囲に】どの楯をも貫き通す矛がある。**(意図通りの現実的意味「A」)**
そして【同じ時間範囲に】どの矛をも通させない楯はない。**(経験による意図から離れた現実的意味 not「B」)**

(2) 意図通りの現実的意味「A」－＞意図通りの空想的意味「B」を文章から直接感ずる意図通りの空想的意味、「A」－＞not「B」を経験から感ずる現実的意味とみなす。

文章から直接感ずる意図通りの空想的意味とみなす：
【ある時間範囲に】どの楯をも貫き通す矛がある。**(意図通りの現実的意味「A」)**
そして【同じ時間範囲に】どの矛をも通させない楯がある。**(意図通りの空想的意味「B」)**
経験に基づく現実的意味：
【ある時間範囲に】どの楯をも貫き通す矛がある。**(意図通りの現実的意味「A」)**
そして、【同じ時間範囲に】どの矛をも通させない楯はない。**(経験による現実的意味 not「B」)**

(3) 命題「A」が共通であるが為に本来同じでなければならないと感じているこれら２つの命題を区別するために、意味「A」と「B」に関してどのような具体的性質を現実から見落としているかを考える。

空想的意味に関して、太字斜体部のように視点を修正すれば、現実に在り得る：
ある時間範囲にどの楯をも貫き通す矛がある。**(意図通りの現実的意味「A」)**
そして***それとは異なる時間範囲に***どの矛をも通させない楯がある。**(意図とは離れた現実的意味「B‘」)**
経験に基づく現実的意味：
ある時間範囲にどの楯をも貫き通す矛がある。**(意図通りの現実的意味「A」)**
そして**その同じ時間範囲に**どの矛をも通させない楯はない。**(経験による現実的意味 not「B」)**

すなわち、矛と楯とが属する時間範囲が「A」と「B」とで異なるのである。そしてそのことが文章の中で記述されていず、さらに時間範囲に関し異なっていないかのように錯覚させる記述があるために現実の経験との間に対立を感じるのである。

以上で、本来異なると認識すべき事を認識できていないことによって矛盾や対立が起きる事を説明した。一方、この事は同じとみなす事が前提である事は極めて重要である。

すなわち、アキレスと亀の競争において追い抜けない場合、アキレスが亀に追いついてからも同じ競争と考えているから矛盾となるのである。もし、追いついてからは競争をやめ、違う追い越さない歩き方に変えたと考えると矛盾しなくなるのである。例え、表現ではそう主張していなくても。また、矛と楯の場合も、どの楯をも貫く矛の場合と、どの矛をも通させない楯の場合も同じ矛と楯と考えるから対立するのである。もし、違うグループの話と考えた場合には、対立しなくなるのである。すなわち、矛盾や対立は、「同じ」と認識することが前提であり、その上で「本来異なる」と認識すべきところを認識していない事によって生じるのである。したがって、「同じ」と認識していなければ、表現はおかしくても矛盾や対立さえ起きないし、矛盾や対立は、「同じ」と認識したうえで「本来異なると認識すべきところ」をそう認識していないことに対する警鐘と考えられるのである。

この事は、人類史を考えてみればよく分かる。すなわち、奴隷制が存在し得たのは、奴隷である人たちを自分と同じ人間とみなさないことによって当然の事として存在し得たと考えられる。それに対して、奴隷を「同じ人間とみなす」（あるいは奴隷自身が「同じ人間だ」と考える）ようになって矛盾や対立が現れ、境遇の相違が意識・克服（すなわち、奴隷からの解放を求めて）され奴隷制の廃止に至ったと考えられるのである。また、一部の人間、すなわち、国王に主権があり、他の人たちに主権がない事を当然の事として、国王の主権による支配下で民が生存する時代があり得たと思われる。それに対して全ての人たちに「同じように主権がある」とみる事が出来るようになって同じく矛盾や対立が現れ境遇の相違が意識・克服（すなわち、自由や平等を求めて）され主権在民の時代になったと考えられるのである。この矛盾や対立が現れてきた状況と境遇の相違が意識・克服される状況とは、日本の明治維新においてより明確に見る事が出来るように思われる。すなわち、幕末から明治維新への変革のエネルギーは寺子屋や藩校等の教育の普及により同じ能力を意識し始めた人たち（特に下級武士）の身分制に対する反発であり、勝海舟、福沢諭吉らによる咸臨丸での渡米、岩倉具視らによる米欧視察などによってはじめて、自由、民権と平等という日本と欧米の体制の違い、あるいは変革の行くつく先を明確に意識し、それ以降その方向へ向かったと考えられる。

4　矛盾、対立の解消と概念分析 [3)]

以上、考察してきたように、人間が感じている意味には、本来、同じとみなすべき事を同じとし、異なるとみなすべき事を異なるとした完全合理的な意味と、そうみなしていない限定合理的な意味とがある事が分かった。ここでは、この限定合理性を完全合理性に近づけてゆくためには概念分析が有効であることを示す。

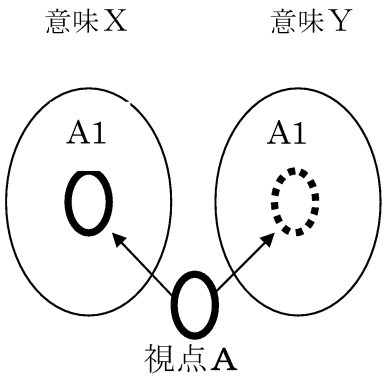
概念分析とは基本的には以下の２つの操作を意味する。

a. ある意味**X**に関して視点Ａから A1 という性質をみつけたら、ある意味**X**に類似した別の意味**Y**に当てはめＡ１に相当する性質を発見する事である。
b. ある意味**X**に対して性質Ａ１が共通していて全体としては意味**X**と異なる別の意味**Z**と比較し、ある視点Ｂから見て異なる性質Ｂ１とＢ２とを発見する事である。

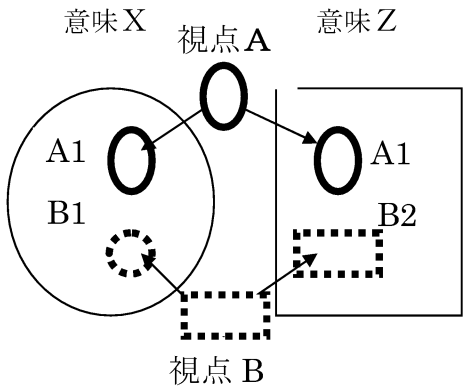
したがって、この操作は基本的には「意味**X**にA1 という性質があるなら、それに類似した別の意味**Y**における A1 に相当する性質は何か」、「意味**X**とそれとは異なる別の意味**Z**においてA1 が共通しているとき、何が異なるのか。」という問題意識を持つ事なのである。この関係を図３に示す。

図３.

(1) 意味Ｘと類似した意味Ｙに対して意味Ｘに視点Ａに相当する性質A1 があるから意味Ｙにも類似した性質A1 があると考える。



(2) 意味Ｘと相違した意味Ｚに対して視点Ａに相当する性質A1 が共通であるから、視点Ｂに相当する異なる性質Ｂ１とＢ２があると考える。



アキレスと亀の競争では、意味Ｘがアキレスと亀が競争をして、アキレスがカメを追い抜くという命題であり、意味Ｚがアキレスと亀が競争をして、アキレスがカメを追い抜けないという命題に相当し、両命題は異なる。それに対して、競争という視点Ａから見てアキレスと亀が競争している事は同じＡ１である。そこで、共通の性質Ａ１を持つのに、両命題ＸとＺが異なるからには異なる性質がある筈、それは何かという問題意識を持つ事によって、その結果、歩幅という視点Ｂから見て、命題Ｘでは「歩幅を縮めない」という性質Ｂ１を持ち、命題Ｚでは「アキレスがカメに追いついてから歩幅を縮める」という異なる性質Ｂ２を持っている事が分かる。

また、矛と盾では、意味Ｘがどの盾をも貫き通す矛があって、どの矛をも貫き通させない盾はないという命題であり、意味Ｚがどの盾をも貫き通す矛とどの矛をも貫き通させない盾があるという命題に相当し両命題は異なる。それに対して、矛と盾という視点Ａから見て、ある時点で存在しているすべての矛と盾を対象にしている事は同じＡ１である。そこで、共通の性質Ａ１を持つのに、両命題ＸとＺが異なるからには異なる性質がある筈、それは何かという問題意識を持つことによって、その結果、存在する時間範囲という視点Ｂから見て、命題Ｘでは「ある時間範囲」という性質Ｂ１を持ち、命題Ｚでは「どの矛をも貫き通させない盾があるのはそれとは異なる時間範囲」という異なる性質Ｂ２を持っている事が分かるのである。この様にして、概念分析によって問題意識を持つことにより、我々の認識の限定合理的状態を完全合理的な状態へと整理してゆく事が出来るのである。

５ まとめ

客観的世界には、本来、人間がある観点から見て、同じ、あるいは、異なるとして認識しなければならない区別がある事を示し、それを完全合理性とした。そして、それは、同じ時点に存在する２つの対象、あるいは、１つの対象とそれが経時変化を経た後の対象との間で成立しなければならないことを明らかにした。そして、経時変化をする前と後の対象が異なる場合は、外部からの作用によってその対象の部分が異なることを明らかにした。さらに、同じ時点の異なる２つの対象、外部から作用を受ける前後の異なる対象との間では、作用と部分の相違との関係、部分の相違と全体の相違との関係は、一対一対応であることを明らかにした。

一方、人間の認識は不完全であり、本来同じとみなすべき２つの性質を同じとみなしていない、あるいは区別していなければならない異なる性質を区別し得ていないことがあることを示し、同じとみなしていない場合は矛盾や対立さえ生じない、そして、同じとみなしていても異なるとみなすべきところをそうみなしていない場合は矛盾や対立となって現れる事を明らかにした。

そして、このような矛盾、対立を解消するためには、問題意識を持つことが必要で、その事によって本来異なる性質を区別して認識出来るようになり、矛盾や対立を解消出来る事を示した。そして、このような問題意識を自ら作り出す手法を概念分析とした。

さらに、実際に重要なのは、このような作為的な文章によって作られる矛盾や対立ではなく、現実に新しい視点が必要となってきている時にその視点を取り入れていない事によって生ずる矛盾や対立とその解消と考えられる。例えば、環境保全が必要となってきた時に、それを取り入れていない産業活動と環境保全は対立するが、環境保全を考慮した産業活動へと変

化する事によって、環境保全とは対立しなくなる[4]。

最後に、「本来、同じとみなすべき、異なるとみなすべき」とは何かという点について簡単に触れておきたい。これは、そう認識して現実に働きかけた結果、何らかの有意な効果を得る事が出来る物の見方といっていいのではないかと思う。

謝辞

定年となりシニヤスタッフとして研究支援にまわりながらこの研究を行うためには、私の所属する第2事業所 研究業務推進室の理解と協力が必要であった。ここに岡野室長、池津室長代理始め業務室の職員、シニヤスタッフ、秘書の皆さんに心から御礼申し上げます。

参考文献

[1]根岸正光：会長就任にあたって，『情報知識学会誌』，Vol.18，No.3，pp.219，2008.
[2]安平哲太郎：「概念分析」，初版，文芸社，pp.39-pp.51，2004.
[3] 安平哲太郎：「概念分析」，初版，文芸社，pp.67-pp.83，2004.
[4]安平哲太郎他：「歴史から洞察される判断基準としての観点」，情報知識学会誌，Vol.16，No.2，pp.75，2006.

情報過程の諸相
Aspects of Information Process

棚次 奎介
Keisuke TANATSUGU
(元)北九州市立大学
(Former) The University of Kitakyusyu
E-mail: tana@kitakyu-u.ac.jp

システムによる制御過程を作用過程と情報過程の統合過程とみなし，実践過程を制御過程と高次情報過程の統合過程として捉えた．特に，情報・知識を対象とする実践過程をとりあげ，高等生物の情報活動のモデルとして内向実践過程を提起した．

The control process by the system was considered to be an integrated process of the action process and the information process, and the practice process was caught as an integrated process of the control process and the higher-order information process. Especially, the practice process intended for information and knowledge was taken up, and the introverted practice process was shown as a model of the intelligence activity of the higher organism.

キーワード:システム，制御過程，情報過程，実践過程
system, control process, information process, practice process

1 はじめに

「第三の波」[1]において，産業社会を超克する近未来社会として描かれた情報（化）社会の到来が言われて久しい．実際，情報技術の進展と相まってコンピュータとそのネットワークが急速に普及し，人々の間に大量の情報が流通する時代が訪れている．

確かに「情報」は，人間にとって，物質・エネルギーと並ぶ重要な資源であり，その取り扱いに関する技術としての情報通信技術の適用は社会に特別な影響を及ぼすものと考えられる．

携帯電話，インターネットの利用が個人の生活スタイルと人間関係，個人と社会との関係などに目新しい変化をもたらしていると指摘する人も少なくない．

そもそも「情報」とは何なのか．多くの研究者がいくつもの答を用意している[2][3][4][5][6][7][8]ものの，未だ情報概念が一つに収斂しているとはいい難い．

本稿は，「情報」そのものを直接取り上げるのではなく，情報が活用される過程に焦点をあてることによって情報のあり様と意義を探ろうとするものである．

論旨の展開は次の通りである．

(1) システムの定義を，情報を利用可能な生物の抽象概念として拡張し，システムが働く過程を制御過程とした．更にスパイラル過程をもつ制御過程を高次の制御過程として位置づけた．

(2) システムが為す高次の働きかけの過程を実践過程と呼び，実践過程を高次制御過程と情報過程（知識過程）の統合過程として捉えた．

(3) 実践の対象が情報や知識である場合の実践過程は全体が情報過程だけで構成されることを示し，人間本来の発展・創造型思考のモデルとして，内向実践過程を提起した．

2 システムと働きかけの過程

2.1 システムとは

「システム」はどの辞書においても類似の解説がなされている．広辞苑では「複数の要素が有機的に関係しあい，全体としてまとまった機能を発揮している要素の集合体．組織．系統．仕組み．」とされ

ている．しかし，これでは，抽象に過ぎてシステムの範囲を確定できず，情報の考察への契機とはなり得ない．本稿では次のように妥当と思われる説明並びにコメントをつけて，辞書の解説を補うこととする．

(1)「まとまった機能を発揮する」とは「システムとしての働きをする」ことと同義である．

(2) システムの働きかけの過程を $y = f(x)$ と書くとして，一般に次のことがらが過程の与件となる．

①対象 x の特定

②目標 y の設定

③方法 f の策定

(3)一般に，方法 f は $\langle x, y \rangle$ に依存して策定される．

(4)システムとはいえ，いつも働いているとは限らないし，動作手順がいつも明確であるとは限らない．刺激に対する反応行為としてのみ動作するシステムも存在する．

(5) (2)によると，システムの働きかけの過程は，本質的に情報過程を内在化しなければならない．

(6)システムは次の２種類に限られる．

①要素として情報機器が組み込まれているようなシステム

②情報過程を内在できる生物または部分システム及び超システム

(7)専ら物理的法則に従って運動する天体は，それがいかに組織的に見えてもシステムというべきではない．

(8)自己内部への働きかけが可能なシステムも存在する．

(9)コンピュータやロボットなどの機械だけでなく，人間をその構成要素とする家庭，企業，国，社会などもシステムといえる．

(10)個々のシステムがどのような機能を有し，どのような状況でそれらを発揮しているのかは様々である．隠れること，逃れることに終始している弱小動物も環境への「働きかけ」を通して賢明に身の安全を確保(ホメオスタシスの維持) している様子が観察される．

(11)「働きかけ」，「目標」，「方法」などは，人間をはじめとする生命体との関係において意味を持つ．コンピュータなどの機械も人間が使用している状況においてはじめてシステム足り得る．したがって，原理的に，生命体が関与しないシステムは存在しないといえる．

2.2 制御システム

システムが目標をもち，それを実現するためには，対象に的確な働きかけを行う必要がある．そうした目標の実現機構を組み込んでいるシステムを制御システムという．

2.2.1 フィードバック制御

制御対象が複雑になると，一般に目標の実現が難しくなる．そこで，目標がどの程度達成されたかを評価して，再度働きかけることが必要となる．そうしたシステムの制御過程を実現する標準的な機構をフィードバック制御(図１)という．

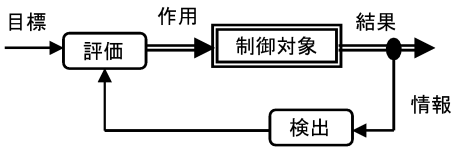


図1　フィードバック制御

図１での用語の意味は次の通りである．

・目標：所与の固定的目標
・評価：目標と検出結果との差異に基づく新たな作用内容の決定
・作用：制御対象への作用
・結果：作用後の制御対象の状態
・検出：作用結果の検知

評価はコンピュータまたは人間によって担われる．

いずれにしろ，制御過程の進行と同時

に内在する情報過程が進行していく．制御過程からこの情報過程を除いた過程を作用過程とすれば，元の制御過程は作用過程と情報過程の統合過程として捉え直すことができる．

フィードバック制御は，図１の各ステップを図２のように繰り返し実行することで初期の目標を達成するスパイラル過程として，定式化される．

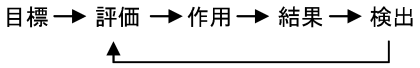


図2　フィードバック制御によるスパイラル過程

2.2.2 シーケンス制御

概念的に最も単純な制御システムはシーケンス制御システム（図３）といわれるもので，工場における自動生産システムや，情報の流れを制御して逐次的に処理を進めるコンピュータがその代表例である．

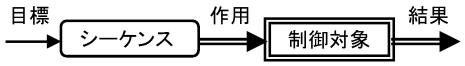


図3　シーケンス制御

フィードバック制御が改善され，スパイラルサイクルを経なくても一様に目標が達成できるようになれば，それはシーケンス制御といえる．

例えば，人間が自転車に乗る練習を始めたとしよう．スパイラル過程はそのまま学習過程でもある．運転方法を習得するまでは失敗の連続であるが，要領を習得すれば，スパイラル過程が終了して，無意識のうちに自転車を乗り回せるようになる．ここの「要領」がシーケンス制御のシーケンスとなる．

2.2.3 PDCA 方式

図４はマネジメントサイクルとしてのスパイラル過程である．

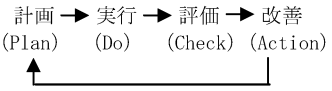


図4　PDCA 方式によるスパイラル過程

フィードバック制御（FBC）によるスパイラル過程と PDCA 方式によるスパイラル過程の間に，図５のような対応関係を見出すことができる．

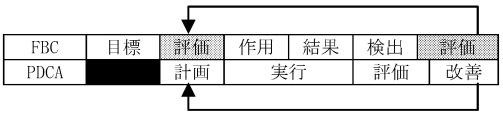


図5　FBC と PDCA の対応表

ここで注意したいのは，フィードバック制御によるスパイラル過程では目標が固定されていることである．

生命体がその生を維持するためには生存環境に適応することが至上の命題であり，大抵の生物は制御対象や環境の変化に対応して目標や働きかけ方を随意に変更することができる．

更に人間の場合は，論理を駆使して，幾多の発明・発見，技術の開発・生産，社会の形成・発展を成し遂げてきた．人間と他の生物との根本的な違いは論理的判断力と制御能力にあると考える．

以降は，PDCA 方式も含めて，スパイラルサイクルを伴う高次の制御過程を考察の対象とする．

2.3 実践過程

高等生物が，目標をもって意識的に対象に働きかけることを実践という．

実践過程はシステムが主体となって遂行される目標実現のための過程であり，その意味では高次制御過程そのものであるが，併せて，認識の検証と深化を伴う高次情報過程（知識過程）が連動する．

実践過程は表１に示すような高次制御過程と情報過程の統合過程のことである．

あらゆる人間活動がこうした実践過程を通じて豊かに展開されていく．

表1　実践過程の構造

実践過程	高次情報過程	知識過程
	高次制御過程	情報過程
		作用過程

自転車の運転のように，実践過程を通して習得され，知識として定着した「方

法」はそれ以降，スパイラル過程の渦に流入してくることはない．その関係は図６に示す通りである．

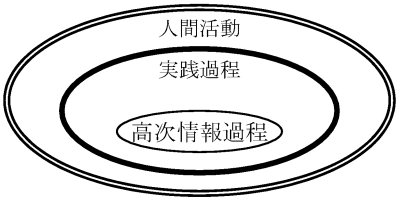


図6　実践過程の位置づけ

次に，実践過程の意味を一層明瞭にするために，その契機と検証に関わる事項を列挙しておく．

＜実践過程の契機＞
(1)目標の設定
(2)制御対象及び環境に関する科学的認識（属性，関係性，法則性など）
(3)制御対象に対する働きかけ方（方策，計画，工程，手順など）の策定
(4)制御対象及び環境の変化に対する対応策

＜実践過程の検証＞
(1)目標は達成されたか
(2)制御対象及び環境に関する科学的認識が深まったか
(3)働きかけ方は的確であったか
(4)予想外の事態に対する対応は的確であったか

実践過程の契機の提供と検証は知識をよりどころとして専ら高次情報過程によって為されるものである．したがって，実践の有効性は，情報過程の実際に大きく依存するものと理解できよう．

2.4 コンピュータは実践できない

コンピュータの情報過程の契機として現れる情報実体と人間の実践過程の契機となる事項を比較すると，データが(2)に,プログラムが(3)に,実行エラーが(4)に対応すると考えられる．しかし，実践過程の契機となる事項には多分に不確かな要素が含まれており，直ちに目標が達成できるわけではない．それに比べてコンピュータの場合はすべての要素が人間によって明確に与えられ，コンピュータはプログラムの指示に従って機械的にデータを処理するだけであり，実践過程に潜むスパイラル過程の本性としての「反省と上昇」の過程は存在しない．

また，プログラムのバグ等によって深刻な実行エラーが発生した場合は，スーパーバイザとしてのＯＳが実行過程に介入してプログラムを強制停止させることも可能である．実行過程としての当該プログラムは何ら解決能力を持たない．

2.5 二つの実践過程

実践過程には外向実践過程と内向実践過程が存在する．

内向実践過程は，人間の知識や記憶内容を対象として行われる実践の過程であり，それ以外の実践過程を外向実践過程という．従来の実践概念に対応する実践過程は外向実践過程に近いものである．

こうした実践概念の拡張によって，あらためて情報過程の重要性とその意義を明らかにすることができる．

現代人は人間活動の一環としてインターネットにアクセスすること，新聞を読むこと，ニュースを聞くこと，本を読むこと,勉強をすること,研究をすること,思案することなどに多くの時間を割いている．これらの情報過程もまたその多くを実践過程として捉えることができる．このような実践的情報過程が内向実践過程である．実践過程は必ず情報過程を伴うことを考えれば，内向実践過程は二重構造の情報過程である．その構造は表 2 に示す通りである．

表 2　内向実践過程の構造

内向実践過程	高次情報過程	情報過程（知識過程）
	高次制御過程	情報過程
		情報過程（知識への作用過程）

実はこの内向実践過程における情報過程の二重性の中に，コンピュータとは異

なる情報処理のメカニズムが隠されていると考えられる.

内向実践過程には，知識への作用過程（第一情報過程),それに連動して遂行される情報過程（第二情報過程)，そしてそれら二つの情報過程を統合した高次制御過程を更なる高みから評価する高次情報過程（第三情報過程）が内在する．この第三の情報過程は知識を拠り所に高次制御過程を統括する過程である.ところが，高次制御過程の結果として知識の変容が生じ得る．スーパーバイザの拠り所の変化は高次制御過程の対象，目標，方法などに影響を及ぼす.

こうした内向実践過程のダイナミズムこそが，自己完結的論理の行き詰まりを回避し，発展，発見，発明を生み出す枠組みを提供していると考えられる.

一般に，外向実践は知識空間と現実世界との関係性を強化する.

内向実践過程は知識空間と現実世界との関係性の弱化を招きがちだが，現実世界が反映された知識空間を参照すること，あるいは現実世界に通用する論理を使用することによって,その強化がなされる.

3 情報過程と情報の特性

これまで,「情報過程」ということばを無定義的に使用してきたが，誤解を残さないためにその説明を加える.

また，情報過程の中で登場する情報の具体的な姿形を示して,「情報」とは何かを解明する道筋の一里塚としたい.

3.1 情報過程

システムの情報過程を考察する.

情報処理を主目的とするコンピュータは，人間の頭脳活動における論理的処理過程をモデルとして作られている．従って両者の情報処理の過程は酷似しており，これらの情報過程を共に次のような各過程の統合過程として示すことができる.

① 記憶過程（取込・書込・保存・取出）
② 処理過程（変換・処理）
③ 伝達過程（発信・転送・受信）

以上の過程のうち，記憶過程と処理過程はシステム内で行われ，伝達過程はシステム内またはシステム間で行われる.

処理過程は記憶過程と連動して遂行される.

実際の入力・取込の過程，出力・提示の過程は，それぞれ記憶過程，伝達過程とみなすことができる.

3.2 情報の基本特性

3.2.1 情報の記憶過程

情報は媒体の状態と対応付けて記憶されるので，一般に媒体の物理的状態を不変に保つことにおいて情報の保存が実現する.

3.2.2 情報の処理過程

一般には，システムの状態の変化が情報の変化に繋がる．したがって，情報処理は状態の変換を通して実現する．電流や電圧の特定の２状態（人間はその２状態を２種の記号０と１で表す）を指定通りに制御できることがコンピュータの根幹をなす仕組みである.

3.2.3 情報の伝達過程

情報の伝達は，ある一定の物理・化学的状態が媒体の中を伝播するか，特定の化学物質が媒体の中を移動することによって実現する．空気の振動パターンが聴覚情報（音の大きさ，高低，音色）であり，光の振動パターンが視覚情報（光の強さ，明るさ，色）である.

3.2.4 情報の物理的特性と独立性

システムの受容機構（感覚器やセンサー）に対して特定の物理･化学的作用（刺激）が発生した場合にはそれをシステムが情報として受理することができる.

例えば，音は空気の振動（可聴周波数：20Hz～ 20kHz）が耳の鼓膜を振動させ，その振動パターンが神経を通じて脳に送

られた情報である．鼓膜が振動してもその振動数や振幅によっては，脳に伝達されない．目で見る像は，網膜に投じられた光線（可視周波数：380nm～780nm）の束が二次元パターンとして視神経を通じて脳に送られた情報である．

動物の感覚器を通して取得される情報は，特定の情報媒体のみを荷体としているが，記号表記による情報は様々な媒体を荷体として，貯蔵あるいは伝達が可能であることから，媒体独立的といえる．感覚器を経由して得られた情報も，記憶の中身を何らかの媒体に表現できれば，媒体独立的となる．すなわち，情報の本質は物質・エネルギーとは独立的に存在しているといえる．

3.3 情報のその他の特性

① 情報は最終的にシステムにおいて受理あるいは取得され，その意味はシステムにおいて解釈され，価値はシステムの働きの中で実現される．

② 情報もシステムの受容機構を経由するまでは単に物理・化学的現象に過ぎない．

③ 中枢神経系を持つ生物はその情報能力により環境への高い適応力を持つ．

④ 全ての生物は遺伝子を持ち，遺伝子情報を活用して自己増殖できるが，中枢神経系では自己の体内細胞が保持する遺伝子情報を制御できない．従って神経系を持っている生物は２種類の独立した情報処理系を有することになる．

⑤ 植物は神経系を持たないので，遺伝子情報，地中からの栄養素の吸収，光合成などで，生物としての恒常性を維持している．

⑥ 遺伝子が織り成す情報過程は，直接的には細胞システムの機能にのみ関与するのだが，間接的には生物個体の機能にも関与し，生命体システムとしてのバランスを見事に編み出している．

⑦ 動物は，エネルギー源の獲得や安全な場所の確保のために移動・行動する．それゆえに状況に応じて的確に行動できなければ生存が危ぶまれる．生物が，環境情報を摂取するための感覚器と，摂取した情報を記憶し処理するための中枢神経系・脳をもつことになったのは一つの必然である．情報の価値はまことに大きい．

⑧ 人手を介さないで動く自動機械は，センサーによって環境情報を取得し，目標値実現に向かって効果的に動作する．

⑨ 物質の有用性はその諸属性と連関の中に見出すことができる．情報の価値性はその意味の中に見出すことができる．

⑩ 情報は外界からの刺激によるものと，体内からの刺激によるものとがあり，最終的には脳内処理を通して情報としての意味をもつ．外界の情報は感覚（視覚，聴覚，味覚，嗅覚，触覚など）器を通して取得される．身体各部の痛みなどの情報は各部位に張り巡らされた末梢神経を通して脳に伝えられる．

参考文献

[1] トフラー，A（徳岡孝夫監訳）：「第三の波」中公文庫，588p.，1982.

[2] 正村俊之：「情報空間論」，勁草書房，357p.，2000.

[3] 吉田民人：「情報と自己組織性の理論」，東京大学出版会，295p.，1990.

[4] 村田晴夫：「情報とシステムの哲学」，文眞堂，227p.，1990.

[5] 西垣通：「こころの情報学」，ちくま書房，236p.，1999.

[6] N. ウィナー：「人間機械論－サイバネティックスと社会」，みすず書房，228p.，1954.

[7] 北川敏男：「情報科学の視座－新しい科学像の探求」，共立出版，282p.，1970.

[8] フォン・ベルタランフィ：「一般システム理論」，みすず書房，288p.，1973.

思考展開図を用いたSQL文の可視化手法

Graphical Expression of SQL Statements using Clamshell Diagram

村川猛彦[1*]，中川優[1]
Takehiko MURAKAWA[1*], Masaru NAKAGAWA[1]

[*1] 和歌山大学システム工学部 (〒640-8510 和歌山市栄谷 930)
Faculty of Systems Engineering, Wakayama University

思考展開図は，一つの図で，課題の詳細と，課題と解決策の関係を容易に知ることができる図式であり，機械設計における発想支援ツールとして発展してきた．本稿では，思考展開図を SQL に適用して図解化・可視化する方式を提案する．その表現にあたり，左半分に SQL 文を細分化して木構造として配置し，それと対応するよう右半分にノード等を配置してそれぞれの意味を記述する．具体的な構成法，SQL 文と図の具体例，および SQL 文の作成・理解のためのシステム構想について述べる．

Thinking process development diagram is a graphical expression from which readers can easily find not only the hierarchy of a given problem but the relationship between the problem and the solution. Although that has been developed as an idea creation support tool in the field of mechanical design, we referred to the restricted version as clamshell diagram to attempt to apply to other fields. In this paper we propose the framework for drawing the diagram of the SQL statement. The basic idea is to supply the hierarchical code fragments of a given SQL statement in the left side of the diagram and to put the meaning written in a natural language in the right. After describing the formulation, we show a few examples of SQL statements and the corresponding diagrams. Moreover we discuss a support system, namely a database of SQL statements, for understanding and writing down the statements.

キーワード: 情報の表現, 思考展開図, SQL, ソフトウェア管理
information representation, thinking process development diagram, SQL, software management

1 Introduction

In the Internet services such as electronic commerce and ticket reservation, it is typical that one or more servers employ some sort of database management system (DBMS) to hold the key data, while Web servers act as the wicket for responding users' requests. Since the communication between Web servers and the DBMS using SQL is done within the server side, the users enjoy the service without regard to the database.

When constructing such a practical database system, the lightweight language like PHP (Hypertext Preprocessor) takes part of the logic and the Web servers' behavior, but the SQL statement is often described as a string in the script le. That is why SQL is considered less serious from the viewpoint of source code maintenance.

Object/Relational (O/R) mapping is a mechanism to enable the developers to write the directions about record retrieval and manipulation. ActiveRecord used together with Ruby on Rails is the driving force. However we have to recognize that O/R mappers only supply the convenience of coding but the optimization of the SQL queries is untouched. For example, ActiveRecord leaves the method for executing any SQL statement directly, from which we can infer that it is a human work that makes the most efficient database access via SQL.

We investigated the application of thinking process development diagrams[1] (TPDDs) to lower process of software development. Although TPDDs have been developed for supporting mechanical design, the originators made the point that they are also useful in planning or organizing something in wider elds and in analyzing the phenomena or the existing products in detail. The authors agree to their assertion, and have been looking into the graphical expression for program understanding.

In this paper we propose the formalism for drawing the diagram of the SQL statement. After introducing the outline of symmetric clamshell diagram, we show the framework together with a few examples of SQL statements and the corresponding diagrams. Moreover we discuss a support system, namely a database of SQL statements, for understanding and writing down the statements in an appropriate way.

2 Clamshell Diagram

Thinking process development diagram is an idea generation support tool. In comparison with other tools such as KJ method, mind map, shborn diagram (cause effect diagram) and TRIZ, TPDD is much more readable since you can nd the whole-part relationships and the problem-solution relationship on a chart. So far as the authors know, idea generation supporting methods have been applied to the upper process of software design, development and improvement. Our approach is an attempt to assist the lower process.

There exist several variants of TPDD, but those diagrams were proposed mainly in the practical viewpoint whereas the structure was not investigated sufficiently. Therefore we focused on a simple version of TPDD and de ned the data

structure to name it symmetric clamshell diagram (SCD) after its shape[2]. To sum up, an SCD consists of a pair of trees arranged axisymmetrically, where the roots of two trees are the both ends while each pair of corresponding leaves are connected in the middle. The left and the right tree means the problem and the solution respectively. Each node has a label containing some message, which is more general at the edges and more speci c in the center. Although the number of level of a tree is unbounded, exactly 3 levels (6-layered diagram) seem to be convenient on screen.

We give a formal de nition of the structure of SCD. An SCD is an undirected graph $G = (V, E)$ where V and E are the sets of vertices (nodes) and edges. The set V is divided into two sets V_l and V_r while E is split up into E_l, E_r and E_c. Let $G_l = (V_l, E_l)$ and $G_r = (V_r, E_r)$, then they are isomorphic trees. Moreover, each leaf v_l in G_l must be connected to the corresponding leaf $f(v_l)$ in G_r where f is the isomorphic function of V_l into V_r, that is $(v_l, f(v_l)) \in E_c$. Since the two subgraphs G_l and G_r are isomorphic, we can store in a computer the constitution together with the messages of all the nodes using a single tree by addressing every node appropriately. This implies that an SCD could be described by an XML document.

After investigating the structure of SCD, we drew the diagrams for a couple of pieces of open source software written in C including GNU Hello[2, 3]. The right half of the diagram shows a tree in which the labels are directory names, le names or code segments except comments. The left tree has the inverted connections so that the diagram may be line-symmetric, and the message on some nodes are the comments extracted from the les and the others are empty. The right-and-left allocation is due to a design concept of TPDD, or the belief that his or her thought should be expanded in terms of "what to do" and then converged by making clear "how to do" and integrating them. Using the drawn diagrams, we assessed the sufficiency of comments to make sure that the de nition of an important function has more ful lling comments.

3 Applying Clamshell Diagram to SQL Statement

Students of the authors' department study SQL after mastering C and learning Java. Some of them have a resistance to this programming language since, say, the syntax is unprecedented for them. One of the goals in the database classes is that the students learn to write down the right queries by themselves.

Typically an SQL query consists of dozens of or a few hundred of bytes, much smaller than a source file written in popular procedural languages such as C. As far as the authors see, there is not a courteous comment attached to an SQL statement in practical system development, though those could be seen in answer books (for example, [4]) or for educational purpose. However it is not hard to partition a given SQL statement into the components (for example, FROM clause and WHERE clause in SELECT statement) and comment on the parts manually. Furthermore, it seems to be promising approach to automatically annotate such code segments using a number of pairs of the code segment and the comment stored in a database. That is how we are now considering the application of SQL statements to clamshell diagrams.

We will describe how to convert a given SQL statement into an SCD chart. In contrast to the way of drawing a conventional TPDD, our formalism arranges the code segments on the left half and the corresponding comments on the right. We adopt the reverse arrangement since the tree expression of an SQL statement on the left comes more naturally to SQL programmers who find the desired query by sending ones many times. Although out policy goes against the traditional way of TPDD, the author believe that this placement gains an advantage in a practical sense.

Imagine that you have a database of well-known jobs and their salaries, namely the table `joblist` which has the attributes `title` and `salary`. If you would like to know the jobs whose income is 50,000 (The currency unit is ignored.) or more, then execute "`SELECT title FROM joblist WHERE salary >= 50000;`". We describe the diagram in Fig. 1. The provided statement is divided into the type of statement "SELECT", the target attribute "title", the FROM clause and the WHERE clause. The query is so simple that the labels are line-symmetric as well as the graph configuration. We consider that the values just after SELECT separated by commas should be expressed as the discrete nodes but complicated condition using AND or OR in the WHERE clause should be single since it is hard to analyze and segmentalize the expression.

Another example is a query including a subquery. To know the jobs which the worker gains the highest income in the list, "`SELECT title FROM joblist WHERE salary = (SELECT MAX(salary) FROM joblist);`" is the code. Figure 2 shows the SCD expression. Note that the subquery forms an SCD.

The variable x in Fig. 2 is a parameter for grafting. When restoring the query and the description, we assign the one constituted by the subtree to x. The right

half of the diagram produces the direction “get the title from the joblist where the salary is equal to the maximum of the salary from the joblist”, where the label “get” on the root of the subquery is removed. What is substituted for each x is unique on this diagram, but if there exists a message with more than one parameter, then the variable name should be speci ed along with the link.

Note that the diagram is about data expression or data structure, and we will have to choose more intuitive indication means.

4 Database of SQL Statements and SCD

To register and maintain a wide variety of SQL statements, the database is the cornerstone. In this section we are discussing the database for SQL statements. The issue includes (1) from what to collect and (2) how to store.

Collecting and selecting the actual SQL statements holds the key to the success of the constructing code repository. We are making a collection of them not only from books but from our development of database applications as well as the classes of database.

To draw an SCD from a given SQL statement, it would be impossible to nd comment of the statement except for written in a book. We will annotate the code segments by hand initially, and subsequently implement the feature of the automatic comment. When each statement is moderately divided, a code segment is stored in the database together with the corresponding comment. In addition, the data for composing an SCD including the references to the code segments is preserved.

Many books for writing SQL statements have been published but the readers cannot invoke the code to verify it. We attempt to construct a database of veri able SQL statements together with execution environments and data set to which the query is applied. Since the learners can see the result quickly, they will be able to know not only the syntax of SQL but good or bad SQL statements in a practical sense efficiently. It may be useful to those who brush up on SQL as well. Finally the database of SQL statements will be informative and helpful to the programmers of database manipulation from beginners to experts.

5 Conclusions

In this paper we have proposed the framework for drawing symmetric clamshell diagram provided an SQL statement. The con guration of the diagram with a pair of trees helps one to understand the whole and the parts of the query. In addition, the manipulation of the tree structure

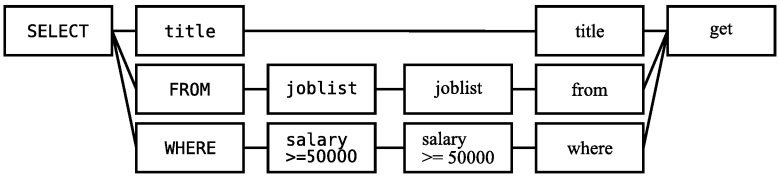


Fig. 1　Clamshell diagram of simple SQL statement

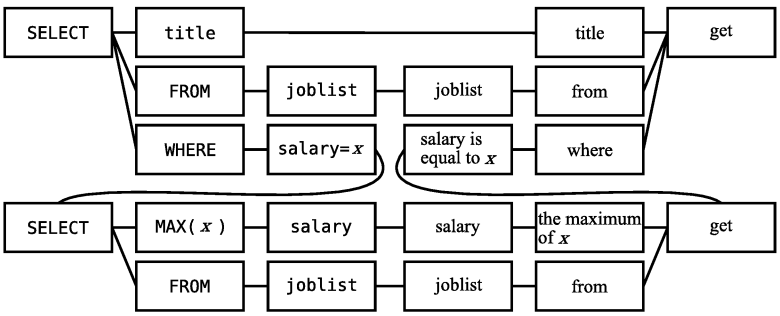


Fig. 2　Clamshell diagram of SQL statement including subquery

such as substitution or grafting will produce another statement.

Future works include expressing SQL codes other than SELECT statements graphically and constructing a practical database of SQL statements together with their SCD con guration.

References

[1] Mase, Hisao; Kinukawa, Hiroshi; Morii, Hiroshi; Nakao Masayuki; Hatamura, Yotaro: “Mechanical Design Support System Based on Thinking Process Development Diagram”, Transactions of the Japanese Society for Arti cial Intelligence, Vol.17, pp.94–103, 2002 (in Japanese).

[2] Murakawa, Takehiko; Kawasaki, Takayoshi; Mizuochi, Hirofumi; Nakagawa, Masaru: “Formulation of Clamshell Diagram and Its Application to Source Code Reading”, Proceedings of the Eighth Joint Conference on Knowledge-Based Software Engineering (JCKBSE 2008), pp.474-483, 2008.

[3] Kawasaki, Takayoshi; Murakawa, Takehiko: “Applying Thinking Process Development Diagram to Source Code Reading Support”, FIT2008, Vol.1, pp.117-118, 2008 (in Japanese).

[4] Celko, Joe: “Joe Celko’s SQL for Smarties: Advanced SQL Programming, Third Edition”, Morgan Kaufmann, 840p., 2005.

サーチエンジン検索結果ページにおける視線情報の分析
Analysis of User Eye Movements during Viewing of Search Engine Results Pages in Web Information Seeking Tasks

高久雅生 [1], 江草由佳[2], 寺井仁[3], 齋藤ひとみ[4],
三輪眞木子[5], 神門典子[6]
Masao TAKAKU [1], Yuka EGUSA[2], Hitoshi TERAI[3], Hitomi SAITO[4],
Makiko MIWA[5], Noriko KANDO[6]

*1 物質・材料研究機構 科学情報室
National Institute for Materials Science
E-mail: TAKAKU.Masao@nims.go.jp
2 国立教育政策研究所 教育研究情報センター
National Institute for Educational Policy Research
E-mail: yuka@nier.go.jp
3 東京電機大学 情報環境学部 情報環境学科
Tokyo Denki University
E-mail: terai@sie.dendai.ac.jp
4 愛知教育大学 教育学部
Aichi University of Education
E-mail: hsaito@auecc.aichi-edu.ac.jp
5 放送大学 ICT 活用・遠隔教育センター
The Open University of Japan
E-mail: miwamaki@code.u-air.ac.jp
6 国立情報学研究所 情報社会相関研究系 / 総合研究大学院大学
National Institute of Informatics / Graduate University for Advanced Studies
E-mail: kando@nii.ac.jp

Web 情報探索行動中のサーチエンジン検索結果一覧ページ（以下，SERP と呼ぶ）に対する行動に着目し，ユーザ実験の方法論により，視線データ，ブラウザログ，事後インタビュー等の情報を包括的に用いて，ユーザ属性，タスク属性，クエリ属性の 3 つの要因と，眼球運動による視線データとの関連を探った．分析の結果，SERP における行動の説明変数としては，他の要因に比べて Information/Navigational クエリの違いによる効果が最も大きく，SERP への遷移の直接的な原因となるクエリ種別により，ユーザの行動を予測しうる可能性が示唆された．

We used an eye tracking tool to investigate the effect of user-based, task-based, and query-based properties when people performed information seeking behaviors on the Web on the search engine results pages (SERP). For the analysis of the user experimental data, we integrated information from various sources, including eye tracking data, browser logs, and post-interview data. An analysis of the results indicated that the query type has a dominant effect and implies that SERPs behaviors might be predictable from the type of query.

キーワード: 視線情報, 検索結果ページ, 情報探索行動
Keywords: Eye movement data, Search engine results pages, Information seeking behavior

1　はじめに

Google に代表されるサーチエンジンが日常的に使われるようになり，検索エンジンの性能向上と探索行動の支援が，より重要となってきている．一般の利用者がサーチエンジンをどのように使っているかは，ログ分析やユーザ実験，インタビュー調査などを通じた研究が行われている [1].

しかしながら，情報探索の過程で情報をいかにピックアップし，その結果として次にどのような行動が選択されるのかについて，ユーザ側での自然な情報探索行動全体にかかわる要因との関係について，包括的に検討している研究は数少ない．これは，情報のピックアップを定量的にとらえる方法が限られていることが挙げられる．

一方，認知科学における問題解決研究や，ヒューマンコンピュータインタラクションの研究分野においては，眼球運動データを用いた分析手法を使って，環境から提供される外的な資源のうち，人がどのような情報に注目し，またそれらの情報を課題の解決にどのように利用しているのかについての研究 [2][3] がある．このような視線データは，ユーザが注目する視点をとらえる生理的側面からのフィードバックであり，近年には Web ユーザビリティ調査などでの利用も一般的なものとなっており [4]，情報検索研究においても検索アルゴリズムへのフィードバック [5][6] といった応用例はあるものの，情報探索行動全体に関わる要因との関連を包括的にとらえた研究は見当たらない．本稿では，この眼球運動データの分析を通じて情報探索行動の精緻な理解を図ることを目指す．

眼球運動データを用いて，検索エンジンを対象とした情報検索における眼球運動の分析についての研究 [7][8][9][5][10] があるが，これらは，検索エンジンを直接的に改良することを目的としたものが多く，眼球運動の分析や考察においても検索エンジンの利用部分に特化したものとなっている．対して本研究ではユーザ実験の方法論に即しながらも，検索エンジンを利用した検索だけでなく，検索結果の先にある Web ページでの利用者の探索行動をも対象とし，できるだけ自然な状況での探索行動を観察することによって，より豊かなユーザ行動の理解を得ることを目的とする．

Broder[11] は，実際のサーチエンジン利用者調査や Web サーチエンジンのクエリログから，サーチエンジンの利用要求を Informational, Navigational, Transactional の 3 つに分類し，その割合などを調査している．この 3 つの分類はそれぞれ，Informational は Web ページの閲覧により情報の獲得を意図した情報要求，Navigational は特定のサイトへの移動を意図した情報要求，Transactional はウェブを介した様々なサービスの遂行を意図した情報要求と定義されている．

Broder によるクエリ種別がサーチエンジンに対する情報要求に限られているのに対し，情報探索研究の文脈では，日常的な探索行動における個別具体的な目的をもったひとかたまりののタスクとして，作業タスク概念の重要性も指摘されている [12]．たとえば，Kellar らは日常的な探索単位として，事実発見，情報収集，ブラウズ，モニター，トランザクションという 5 つのタスク種別を定義している [13].

著者らは，ユーザ実験の方法論に基づいて，2 つの作業タスクを探索課題として与えたうえで，習熟度などの点から異なると思われるユーザ層を対象としたユーザ実験を行い，事前アンケート，検索遂行中の行動データ，発話，視線データ，および事後イン

タビューといった複数の情報を収集し，分析してきた [14][15][16][17]．本研究では，課題遂行中におこなわれた検索行動から得られたクエリに対して Broder の定義にもとづく分類を行い，分析した．これらの分析を通じて，探索行動から得られる定量的な情報としての視線データに対して，タスク要因，クエリ要因，さらに検索者の属性などとの関連を分析し，情報探索行動の精緻な様相を明らかにすることを目指す．本稿では，探索遂行中のサーチエンジン検索結果ページを対象とした視線データに関する分析結果を報告する．

1.1　仮説

本研究では特に，ユーザによる自然な情報探索行動の過程を観察することにより，情報探索プロセスにかかわる以下の要因を挙げ，それぞれの要因間の関係を明らかにすることを目指す．

1. ページ属性（情報を獲得する閲覧サイト，閲覧ページに関する情報）
2. システム属性（サーチエンジンの種類，検索アルゴリズムなど）
3. ユーザ属性（Web 情報探索への習熟度，検索スキルなど）
4. タスク属性（タスクの目的，種類など）
5. クエリ属性（検索システム等に投入するクエリの種類など）

ただし，本稿での分析対象がサーチエンジンの検索結果一覧ページ（以下，SERP と呼ぶ）における視線データに限られるため，ページ属性を今回の分析の対象から除く．さらに検索システムに関しても，後述するユーザ実験において観測された，Google, Yahoo! Japan, MSN の 3 つの一般的なサーチエンジンのみを対象として分析することから，システム属性についても今回の分析の対象から除く．

よって以下では，ユーザ属性，タスク属性，クエリ属性の違いが視線データに与える影響について明らかにする．

2　ユーザ実験

2.1　実験参加者

学部生 11 名（男性 5 名，女性 5 名，平均年齢 20.0），大学院生 5 名（男性 4 名，女性 1 名，平均年齢 24.6）が実験に参加した．学部生の専攻は，経済学，文学，工学，外国語（スペイン語学科），心理学，理学，土木工学など多様であった．また，大学院生は参加者全員が図書館情報学の専攻であった．

探索実験は 2007 年 11 月（学部生 11 名），2008 年 3 月（大学院生 5 名）の 2 回にわけて実施し，実験参加者には謝金を支払った．

2.2　タスク

被験者は，レポート課題と旅行課題の 2 種類の検索課題に取り組んだ．レポート課題は，「世界史」のレポートを書くために必要な情報を探索するという課題であった．被験者にとって親近性のある課題とするため，具体的なテーマは被験者自らが興味のある対象を選んだ．一方，旅行課題は，身近な人との「旅行計画」のために必要な情報を探索するという課題であった．レポート課題と同様に，具体的な目的地等は被験者が興味のある場所を選んだ．

2.3　実験手順

被験者は事前アンケートとして，Web 検索エンジンを用いた日常的な検索経験についての質問に回答した．発話練習を兼ねた練習課題に 5 分間取り組んだ後，2 つの検索課題にそれぞれ 15 分間取り組んだ．検索課題はそれぞれレポート課題と旅行課題で，順序による影響を考慮し，被験者によって課題の遂行順序がランダムになるよう配置した．

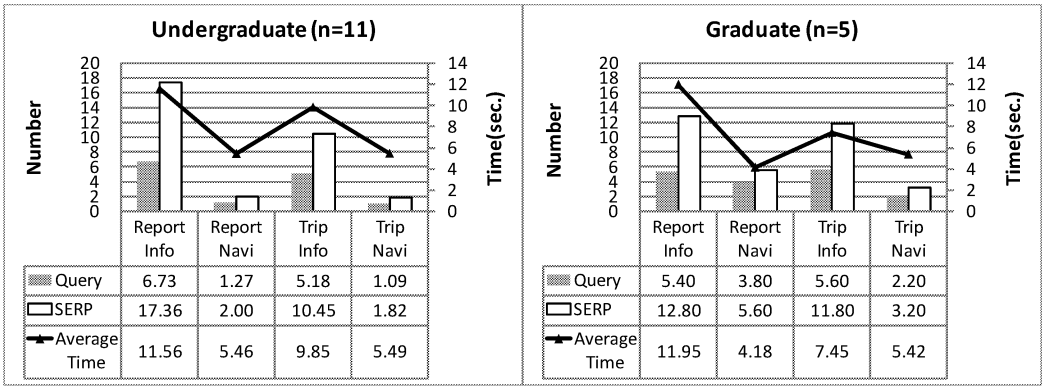


図 1　一課題あたりのクエリ数と SERP へのアクセス

2.4　収集データ

検索遂行中の被験者の眼球の動きは，眼球運動測定装置（Voxer ST-600）によって計測した．実験には 19 インチ液晶モニタに対し，1024 × 786 の解像度に設定した Windows XP の PC を使用した．Web 閲覧用のブラウザには Firefox を使用した．ブラウザのブックマークには Google および Yahoo!Japan のトップページを追加し，日常での自然な探索と同様の状態となるよう，自由に探索を行うよう指示した．なお，視線データの分析のために，原則としてブラウザは画面上にウィンドウサイズを最大化した状態で使用するよう促し，必要に応じてタブ機能を使うよう指示した．また，ブラウザのログは Slogger，コンピュータの画面は画面キャプチャソフト HyperCam によって記録した．また，課題遂行中の思考内容を明らかにするため，被験者には検索遂行中に発話するよう求めた．

各課題終了後には，課題についての困難度や満足度等を問う事後アンケートを実施した．さらにアンケート終了後，課題に取り組んでいる際の被験者の情報探索行動について，インタビューを実施した．インタビューでは検索時の記憶想起を促すため，画面キャプチャ映像を参照しながら進めた．

3　結果

以下では，2 章で述べた実験から獲得したデータの分析結果を報告する．

3.1　クエリ分類

視線データの分析に先立ち，1.1 節で述べたクエリ属性として，課題遂行中にサーチエンジンに発行されたクエリを，Broder[11] の分類にもとづき分類した．1 章で述べたように，Broder[11] によるクエリ分類は，Informational, Navigational, Transactional の 3 つであり，Informational は Web ページに含まれる情報の獲得を意図した情報要求，Navigational は特定のサイトへの移動を意図した情報要求，Transactional はウェブを介した様々なサービスの遂行を意図した情報要求と定義している．本研究では Broder の分類を参考にしたものの，このうち分類が困難な Transactional クエリは採用せず，課題遂行中に発行されたクエリを Informational クエリもしくは Navigational クエリのいずれかに分類した．クエリ分類は著者のうち 4 名の合議によって決定した．Informational および Navigational クエリの例を表 1 に挙げる．表 1 によれば，レポート課題の Informational クエリでは選択し

表 1　Informational クエリ，Navigational クエリの抜粋

	Report task		Trip task	
	Info. query	Navi. query	Info. query	Navi. query
学部生 ($n = 11$)	British East India Company; アメリカ 起源; エリザベート ハプスブルク 系図; ケネディ リンカーン; 朝鮮 鎖国; 第二次世界大戦 被害; 三国志 歴史	ヒトラー wikipedia; 東インド会社 wiki; msn; 清 ウィキペディア; ウィキペディア; 朝日新聞; 東京工業大学	羽田 集合場所; 飛行機; 越後湯沢; 広島 原爆ドーム; 北海道 観光名所; 沖縄 おすすめ観光地; 沖縄; 沖縄 本島; 京都 イベント 花見; 摩周湖	飛行機 ana; ユースホステル; 宝塚; HIS; プリンスホテル 沖縄; 京都 市バス
大学院生 ($n = 5$)	4750311677[1]; 仏教 成立について; ユーゴスラヴィア解体; ハワイ語 歴史; ワイマール ナチス	amazon; CiNii; ハワイ Wiki	東北 イベント; 屋久島 みどころ; 京都 銀閣寺 モデルプラン; 新幹線 料金; 四国 ホテル	酒田市 wiki; 屋久島 JTB; 乗り換え案内; 世羅高原ファームランド; 深夜バス 四国 キラキラ

た世界史のトピックに近いキーワードが見られ，Navigational クエリではそのテーマに関するウィキペディア上の項目や Amazon，CiNii といった別の検索サイトへと飛ぶようなキーワードが観察された．旅行課題においては，Informational クエリでは選択した行き先の地名や見どころを探すようなキーワードとなり，Navigational クエリではその行き先までの交通機関のサイト名や施設名などが現れた．

図 1 に，ユーザ種別，課題種別，クエリ種別毎の SERP へのアクセス状況として，一課題あたりのクエリ数（Query），SERP 閲覧回数（SERP），SERP 閲覧時間の平均（Average Time）を示す．ここでクエリ数は，サーチエンジンへクエリを発行した回数を指す．SERP 閲覧回数には，SERP 中の次の 10 件を表示するためにページ遷移が加わっているほか，戻るボタンなどのブラウザ操作によって再表示した回数も加えている．また，SERP 平均閲覧時間は検索結果ページが表示されてから，別のページに移るまでの経過時間の平均となっている．

図 1 のように得られた SERP に対するアクセス状況に対して，3 要因混合分散分析を行った．クエリ数，SERP 閲覧回数，SERP 閲覧時間の 3 つに関していずれも，クエリ種別要因の主効果のみが有意であった．まずクエリ数においては，Informational クエリの方が Navigational クエリよりも有意に多く発行された（$F(1, 56) = 53.35, p < .01$）．これは課題種別，ユーザ種別を問わず一貫した傾向として観察された．SERP 閲覧回数でも同様の結果となった（$F(1, 56) = 14.15, p < .05$）．また，SERP 平均閲覧時間では，Navigational クエリにより表示された SERP の閲覧時間の方が，Informational クエリの場合よりも有意に短かった（$F(1, 56) = 14.41, p < .01$）．

3.2　眼球運動の分析

眼球運動測定装置から得られた実験参加者の視線データに基づく分析を以下に示す．なお，実験参加者のうち学部生 2 名は，測定装置の環境において視線データが得られなかったため，分析から除外した．

[1] ある参考書籍の ISBN 番号を表わす．

3.2.1　SERP Lookzone の定義

ブラウザ操作

1: タイトルバー，2: メニュー，3: ブックマーク, 4: ツールバー, 5: URL バー，6: 検索バー，7: 検索バーボタン，8: タブ，12: スクロールバー，21: ページ内検索，22: ステータスバー

SERP ページコンテンツ内

9: サービスリンク，10: クエリボックス，11: 検索ボタン，13: ヒット件数，14: スポンサーリンク，15: スペルチェック，16: タイトル，17: スニペット，18: URL, 19: 関連検索，20: 次へ

図 2　SERP Lookzone

ユーザが情報探索を行う画面上のブラウザ操作関連の要素および SERP ページコンテンツ内の共通要素を，図 2 に示す 22 項目の SERP Lookzone として定義し，注視点のタグ付けに使用した [14]．また，図 3 は SERP Lookzone の各ブロックの位置を Google の検索結果ページで示したものである．なおこれらの項目は今回の被験者が使用した他のサーチエンジン（Yahoo!Japan, MSN）でも同様に分類することができる．

次に実験時に記録した被験者の視線情報を元に，SERP の表示が開始された時刻を基点にして 0.5 秒間隔で画像を切り出した．その上で，抽出した画像に表示された注視点が Lookzone のいずれのに当てはまるかを人手で確認しながらタグ付けした．以下での視線分析ではこの 0.5 秒間隔で抽出した注視点に対する分析を行った．

眼球運動測定装置を用いた分析では一般に，眼球運動により観察されたデータが一定の領域に一定時間（150〜500msec 程度）停留した場合に，これを停留点として，当該領域への注視が行われたとすることが多い [18]．本稿での分析ではこれを用いる代わりに，0.5 秒おきの注視点を用いるにとどめた．これは，Web 情報探索行動が非常に迅速な情報の取捨選択をともなうものであり，かつ，眼球運動測定の対象となる画面のブラウザ上でのスクロールやページ遷移行動などによって，刺激提示画面上の X-Y 軸による画一的な分析になじまず，停留点の単位と画面上で素早く変化する注視領域の判定を行うことが困難であるために，このような分析を行った．検索結果ページのみを対象とすれば，停留点にかかわる分析も一定の範囲で可能であるが，自然な状況での情報探索を行うとの本研究の目的に沿わないものとなる懸念から，その方式は採用しなかった．停留点分析には概して専用ソフトウェアが必要であり，かつ，上記で述べたような制約もあるため採用しなかったが，一方で一定時間間隔による注視点に対するタグ付けは人手による労力を伴うものの，わかりやすい Lookzone が定義されてあれば比較的容易であり，高価な専用の分析ソフトウェアを使う必要もなく，安価なビデオツールの組み合わせで分析できる点に利点がある．

以下，上記の注視ブロックとして定義した Lookzone に対して一課題あたりどれくらいの注視があったかを示す．ただし，設定された Lookzone のうち，実際のユーザ実験における注視がほとんどなく，一課題あたりの平均注視回数が 1 回に満たない 10 ブロックは分析から除外し，以下ではそれ以外の 12 ブロックの分析結果を示す．ちなみに，この除外されたブロックとしては，ブラウザ操作領域で 7 つ（「ツールバー」，「ステータスバー」，「メニュー」，「URL バー」，「スクロールバー」,「検索バーボタン」,「ページ内検索」）あり，SERP ページコンテンツ

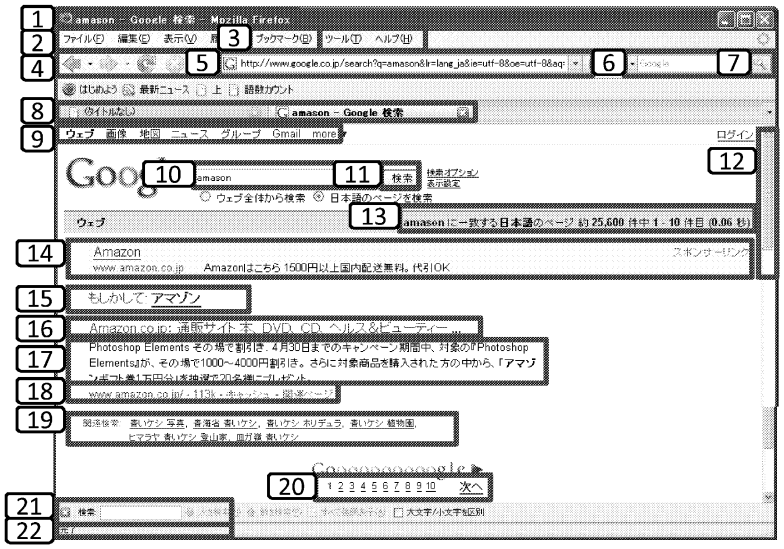


図 3　Google における SERP Lookzone の例

内要素で 3 つ（「次へ」,「ヒット件数」,「スペルチェック」）あった．つまり，たとえば，SERP 閲覧中の「スクロールバー」や「URL バー」への注視はほとんど起きなかったことが分かる．

表 2 に一課題あたりの各 Lookzone の注視回数平均を示す．表 2 からも明らかなように，Lookzone として設定したブロックのうち多くは注視回数そのものが少なかったり，全く注視を受けないブロックも多い．一方で,「タイトル」,「スニペット」,「URL」といった SERP において中心となる検索結果文書の情報に関するブロックに注視が集中していることが分かる．

表 2 に示された各 Lookzone の注視回数に対して，3 要因混合分散分析を行った結果を以下に示す．

ブラウザ操作関連では,「タブ」ブロックにおいて有意差が観察され,「検索バー」において有意傾向が見られた．「タブ」ではクエリ種別の主効果があり，Informational クエリによる SERP の方が Navigational クエリの場合よりも多く「タブ」ブロックを注視していた（$F(1,48) = 6.07, p < .05$）．「検索バー」では，学部生の場合には「検索バー」への注視が全く見られなかった一方で，院生では注視があった．

ブラウザ内の SERP ページ内コンテンツのブロックでは,「スニペット」,「タイトル」,「URL」,「クエリボックス」,「スポンサーリンク」,「関連検索」の 6 領域で主効果が観察された．「スニペット」では，タスク種別とクエリ種別の交互作用に有意差があり（$F(1,48) = 7.61, p < .05$），下位検定の結果，レポート・旅行課題双方において Navigational クエリよりも Informational クエリの場合の方が有意に多くの注視を集め（$F(1,48) = 12.04, p < .01$; $F(1,48) = 4.75, p < .05$），かつ，Informational クエリの場合にはレポート課題の方が旅行課題よりも有意に多くの注視を集めていることが分かった（$F(1,48) = 8.82, p < .05$）．「タイトル」ではクエリの主効果があり，Informational クエリの場合の方が Navigational クエリよりも多くの注視を集めるこ

表 2　一課題あたりの各 Lookzone への注視回数

		Report task		Trip task	
		Info. query	Navi. query	Info. query	Navi. query
学部生	タブ	7.44	0.44	6.44	2.78
($n = 9$)	検索バー	0.00	0.00	0.00	0.00
	タイトルバー	3.67	0.11	0.67	0.33
	ブックマーク	4.20	0.00	0.00	0.00
	スニペット	87.22	1.56	31.00	6.00
	タイトル	57.22	2.22	35.33	6.78
	URL	36.89	0.44	12.44	3.11
	検索ボタン	0.56	0.00	0.67	0.00
	クエリボックス	34.67	0.67	11.00	1.56
	サービスリンク	16.22	1.11	5.00	0.11
	関連検索	1.78	0.11	2.44	0.11
	スポンサーリンク	4.89	0.11	9.89	2.56
院生	タブ	8.80	3.20	5.60	0.40
($n = 5$)	検索バー	3.40	3.00	3.80	0.20
	タイトルバー	0.20	0.20	0.80	0.00
	ブックマーク	0.00	0.00	0.20	0.00
	スニペット	63.80	11.00	26.60	1.80
	タイトル	35.20	6.00	36.60	2.60
	URL	16.00	2.00	11.00	1.40
	検索ボタン	0.00	0.00	0.00	0.20
	クエリボックス	4.00	1.60	2.00	1.00
	サービスリンク	2.20	0.20	1.40	0.80
	関連検索	1.00	0.20	1.00	0.20
	スポンサーリンク	0.00	0.00	11.00	0.40

とが分かった（$F(1, 48) = 19.96, p < .01$）．「URL」においてもクエリの主効果があり，Informational クエリの場合の方が Navigational クエリよりも多くの注視を集めることが分かった（$F(1, 48) = 19.43, p < .01$）．「クエリボックス」でもクエリの主効果があり，Informational クエリの場合に Navigational クエリの時よりも有意に多くの注視を集めた（$F(1, 48) = 5.48, p < .05$）．「スポンサーリンク」においては，タスク種別とクエリ種別双方に主効果が見られ，Informational クエリの場合の方が Navigational クエリの時よりも多くの注視を集め，かつ，旅行課題の場合の方がレポート課題の時よりも多くの注視を集めた（$F(1, 48) = 8.19, p < .05$; $F(1, 48) = 5.74, p < .05$）．「関連検索」ではクエリの主効果があり，Informational クエリの場合の方が Navigational クエリよりも多くの注視を集めた（$F(1, 48) = 7.23, p < .05$）．

3.3　文書ランキング箇所への注視

表 2 にも見られるとおり，SERP においては「スニペット」,「タイトル」,「URL」からなる検索結果文書の要約箇所がユーザの注目をもっとも集める領域となる．つづいて，これらの要約箇所の注視がそれぞれの文書ランク毎に異なるかを分析する．要約

箇所の領域（前述の「タイトル」,「スニペット」,「URL」の 3 ブロック）の注視回数をまとめてランキングごとに分け，どのランクに対する注視が多いのかをクエリ種別ごとに示す．

図 4 は，ランクごとに集計した注視回数をランク 1 位の注視回数を 1 として正規化したものである．Eyemark.info は Informational クエリの時の SERP 内での視点注視回数をあらわし，Eyemark.navi は Navigational クエリにおける注視回数を示す．今回の実験においては，Navigational クエリに対する SERP ではユーザは最高でもランク 10 位の文書への注視に留まった．対して Informational クエリでは最大でランク 50 位の文書要約まで注視された．

4　考察

4.1　クエリ属性

以上に述べてきた分析結果は，SERP における行動の説明変数としては，他の要因に比べて，クエリ種別による効果が最も大きいことを示唆するものであった．このことは SERP がクエリにより駆動されるという性質上，クエリ種別が最も直接的な影響を与えることを考えれば，理解できる．とりわけ，Navigational クエリはサーチエンジン上で特定のサイトの所在を探して，そこへ到達することを目的とすることから，SERP 上でも最上位に表示された文書へとすぐに移っていく傾向が分析結果からも示されている．たとえば，閲覧回数・閲覧時間の短かさや，図 4 に見られるように 1 位もしくは 2 位，3 位へ視線が集中し，以降の下位ランクの文書要約は注視を受けることがまれであるという現象にあらわれている．また，SERP 中の「クエリボックス」への視線注視で違いが出た点も，Navigational タスクでは SERP 上でクエリの修正に時間をかけることなどをしないという傾向を示している．逆に Informational タスクでは．より下位ランクの文書要約までの吟味やクエリのリファインに時間をかけるといった現象が見られた．

4.2　ユーザ属性

クエリ属性とは対照的に，SERP 上の注視回数においてはユーザ属性の影響はほとんどあらわれなかった．「検索バー」,「検索ボタン」といったブラウザ操作機能部分にある注視領域において，ユーザ属性の影響がわずかながらも見られた．これは，一般の大学学部生と図書館情報学専攻の大学院生という属性の違いが Web ブラウザの操作スキルに違いをもたらしたことに起因するものと思われる．事前アンケートの結果によれば，大学院生群の実験参加者は全員が日頃からタブブラウザを使用すると答えており，一方で学部生の参加者では半数以上の実験参加者が Internet Explorer[2] を使用していると回答しており，実際の実験においても大学院生の実験参加者はタブブラウズ機能を多用する傾向があったものの，学部生ではそれほど見られなかった．そういったブラウザ操作に関する習熟度の違いが差異をもたらしたとも考えられる．たとえば，サーチエンジンの利用の際に日頃から検索バーを通じた検索機能を利用していれば，自らのタスク遂行にあたって，わざわざサーチエンジンへページを移らずともその場で検索でき，時間短縮ができると考え

[2] 実験を実施した 2007 年 11 月時点では，タブブラウザ機能を追加した Internet Explorer 7 が公開されていたが，その利用率は 28.9%[19] にとどまっており，大半の参加者はタブブラウザ機能を持たない Internet Explorer 6 以下を使っていたと推測されるが，正確な常用ブラウザのバージョンに関しては確認していないため，不明である．

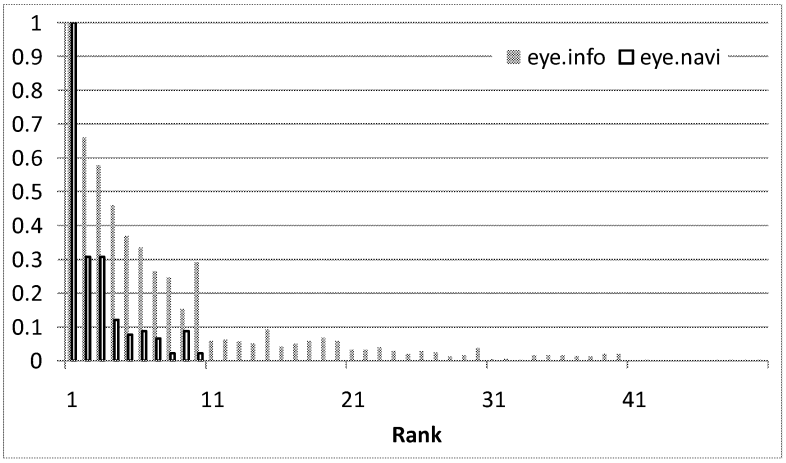


図４　閲覧された文書要約箇所のランク毎の注視回数

たものと思われる．このようなブラウザ環境に対する習熟が影響を与えた部分が見られた一方で，ユーザ属性が持つ検索スキルや Web リソースに対する理解度の違いが与えた箇所については，今回分析した SERP と視線データの範囲では確認できなかった．操作的知識に留まらないより広範なリテラシーを情報探索においてどのように活用するかは重要な研究課題であり，今後，分析の範囲を広げて確認する必要がある．

4.3　タスク属性

タスク属性の影響の度合は，ユーザ属性よりは大きいと思われるものの，予想よりも狭い範囲に留まったと言える．タスク属性としてのレポート課題と旅行課題との違いとして明確に見られたものは，「スニペット」および「スポンサーリンク」への注視回数の差にとどまる．「スニペット」の場合はレポートのための情報収集という性質から，ヒットした文書要約を通じて自らの情報要求に適合するサイトであるかどうかを読み取りながらの探索を行う傾向があったために，旅行課題よりも「スニペット」への注視が集まったと考えられる．旅行というタスクの性質から，「スポンサーリンク」に広告として掲載されるリンク先が重要な情報源となることもありうるとも思われる．また一般に，サーチエンジン上でも多くの「スポンサーリンク」が表示されることも原因の一つと思われる．旅行・観光業界はもともと広告費の多い業種であり，インターネット広告の大手代理店サイバー・コミュニケーションズによれば，2007 年第 4 四半期における交通・レジャー業種の広告売り上げは全体の構成比 8.4% であり，全 17 業種の分類中でも「金融・保険」「情報・通信」「不動産・住宅設備」に次いで 4 番目に売り上げの多い業種である [19, p.59]．ただし，スポンサーリンク注視がこのような出稿量の多さによって生起されたものであるか，むしろユーザにとって欲しい情報が提供された結果によるのかについては，本稿では分析に至っていない．

また，今回設定した 2 課題はいずれも Kellar ら [13] によるタスク種別の定義によれば，情報収集タスクにあたるものと見ることもでき，その点では，タスク自体がさほどの違いを生むような要因となっていないとも考えられる．ただし，以前の分析 [15]

からSERPに限らないWeb全体を使った閲覧パターンでは，旅行課題の方がレポート課題よりも閲覧ページ数が多く，すばやい情報収集をしているなどといった違いがあることが分かっている．このような違いがタスク種別そのものの違いと言ってよいのか，もしくはタスクに特有の探索方略やそのための問題解決空間の構造として現われているのか，とりわけ，これらタスクの特性そのものを検討していくことが今後必要となると思われる．今後の課題としたい．

5　おわりに

本稿ではサーチエンジン上のSERPにおいて，ユーザ属性，タスク属性，クエリ属性の違いという3要因が視線データに与える影響について分析した．SERPに対して直接的な関係となるクエリ属性としてとらえたInformational/Navigationalのクエリ種別がSERPユーザ行動の説明変数として大きな影響を与えていることが分かった．タスク属性およびユーザ属性は比較的少数の変数との関係が見られるにとどまり，SERPおよび視線データに限っては，広範な影響を与える変数ではないことが示唆された．

著者らは，探索型情報検索のためのユーザの情報探索行動を理解するための研究開発に取り組んでいる．情報探索行動にかかわる精緻な分析のために，今後は分析の範囲を広げ，探索のプロセスや探索時のモード切り替えなど，より大きな研究課題との相関を探ることが今後の課題である．

謝辞

本研究は国立情報学研究所公募型研究「情報探索行動の認知モデルの構築とその応用に関する研究」の支援を受けた．

参考文献

[1] Spink, Amanda; Jansen, Bernard J.: “Web Search: Public Searching of the Web”. Kluwer Academic Publishers, Dordrecht, Netherlands, 199p., 2004.

[2] Terai, Hitoshi; Miwa, Kazuhisa: “Sudden and Gradual Processes of Insight Problem Solving: Investigation by Combination of Experiments and Simulations”, *Proceedings of 28th Annual Meeting of the Cognitive Science Society*, pp. 834–839, 2006.

[3] Jacob, Robert J. K.; Kari, Keith S.: “Eye tracking in human–computer interaction and usability research: ready to deliver the promises (section commentary)”. Hyona, Jukka; Radach, Ralph; Deubel, Heiner, editors, *The mind’s eye: cognitive and applied aspects of eye movement research*, pp. 573–605. Elsevier Science, Amsterdam, 2003.

[4] Nielsen, Jakob: “Eyetracking Research into Web Usability”. URL: `http://www.useit.com/eyetracking/`（2009年4月10日参照）.

[5] Granka, Laura A.; Joachims, Thorsten; Gay, Geri: “Eye-tracking analysis of user behavior in WWW search”, *SIGIR ’04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 478–479. ACM, 2004.

[6] Joachims, Thorsten: “Optimizing search engines using clickthrough data”, *Proceedings of KDD 2002*, pp. 133–142. ACM, 2002.

[7] Lorigo, Lori; Haridasan, Maya; Brynjarsdottir, Hronn; Xia, Ling; Joachims, Thorsten; Gay, Geri; Granka, Laura; Pellacini, Fabio; Pan, Bing: “Eye tracking and online search: Lessons learned and challenges ahead”, *JASIS&T*, Vol. 59, No. 7, pp. 1041–1052, 2008.

[8] Lorigo, Lori; Pan, Bing; Hembrooke, Helene; Joachims, Thorsten; Granka, Laura; Gay, Geri: “The influence of task and gender on search and evaluation behavior using Google”, *Inf. Process. Manage.*, Vol. 42, No. 4, pp. 1123–1132, 2006.

[9] Guan, Zhiwei; Cutrell, Edward: “An eye tracking study of the effect of target rank on web search”, *CHI ’07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 417–420, New York, NY, USA, 2007. ACM.

[10] Pan, Bing; Hembrooke, Helene; Joachims, Thorsten; Lorigo, Lori; Gay, Geri; Granka, Laura: “In Google we trust: Users’ decisions on rank, position and relevancy”, *Journal of Computer-Mediated Communication*, Vol. 12, No. 3, pp. 801–823, 2007.

[11] Broder, Andrei: “A Taxonomy of Web Search”, *SIGIR Forum*, Vol. 36, No. 2, pp. 3–10, 2002.

[12] Ingwersen, Peter; Jävelin, Kalervo: “The Turn: Integration of Information Seeking and Retrieval in Context”. Springer, 448p., 2005.

[13] Kellar, Melanie; Watters, Carolyn; Shepherd, Michael: “A field study characterizing Web-based information-seeking tasks”, *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 7, pp. 999–1018, 2007.

[14] 高久雅生; 寺井仁; 江草由佳; 齋藤ひとみ; 三輪眞木子; 神門典子: 「Web 情報探索における視線データの予備的分析」, 情報知識学会誌, Vol. 18, No. 2, pp. 181–188, 2008.

[15] Terai, Hitoshi; Saito, Hitomi; Egusa, Yuka; Takaku, Masao; Miwa, Makiko; Kando, Noriko: “Differences between informational and transactional tasks in information seeking on the web”, *Proceedings of IIiX 2008*, pp. 152–159. ACM, 2008.

[16] 齋藤ひとみ; 江草由佳; 高久雅生; 寺井仁; 三輪眞木子; 神門典子: 「Web 情報探索行動の分析: 課題の志向性と経験の違いによる影響についての予備的検討」, 電子情報通信学会研究報告「Web インテリジェンスとインタラクション」研究会 (IEICE SIG-WI2) 第 13 回研究会, pp. 37–42, 2008.

[17] Egusa, Yuka; Takaku, Masao; Terai, Hitoshi; Saito, Hitomi; Kando, Noriko; Miwa, Makiko: “Visualization of User Eye Movements for Search Result Pages”, *Proceedings of EVIA 2008 (NTCIR-7 Pre-Meeting Workshop)*, pp. 42–46, 2008.

[18] 斎田真也: 「読みと眼球運動」. 『眼球運動の実験心理学』. 第 8 章, pp. 167–197. 名古屋大学出版会, 名古屋, 1993.

[19] 財団法人インターネット協会監修:「インターネット白書 2008」. インプレス R&D, 東京, 368p., 2008.

第 17 回(2009 年度)情報知識学会研究大会実行委員会

実行委員長	研谷 紀夫	東京大学
副実行委員長	村井 源	東京工業大学
	江草 由佳	国立教育政策研究所
	高久 雅生	物質・材料研究機構

情報知識学会誌　編集委員会

編集委員長	国沢 隆	東京理科大学	
副編集委員長	芦野 俊宏	東洋大学	
編集委員			
相田 満	国文学研究資料館	石井 守	情報通信研究機構
石塚 英弘	筑波大学	岩田 覚	東京大学
内田 努	北海道大学	宇陀 則彦	筑波大学
江草 由佳	国立教育政策研究所	大久保 公策	国立遺伝学研究所
岡本 由起子	元東京家政学院大学	小川 恵司	凸版印刷(株)
神立 孝一	創価大学	五島 敏芳	京都大学
阪口 哲男	筑波大学	白鳥 裕	大日本印刷(株)
菅原 秀明	国立遺伝学研究所	太原 育夫	東京理科大学
田良島 哲	東京国立博物館	時実 象一	愛知大学
中川 優	和歌山大学	長田 孝治	東京都ビジネスサービス(株)
長塚 隆	鶴見大学	中山 尭	神奈川大学
中山 伸一	筑波大学	西川 信孝	みずほ情報総研(株)
西澤 正己	国立情報学研究所	西脇 二一	奈良大学
根岸 正光	国立情報学研究所	原 正一郎	京都大学
原田 隆史	慶應義塾大学	藤井 賢一	産業技術総合研究所
藤原 譲	筑波大学名誉教授	細野 公男	慶應義塾大学名誉教授
村川 猛彦	和歌山大学	安永 尚志	人間文化研究機構
山本 毅雄	国立情報学研究所名誉教授	山本 昭	愛知大学

■複写をされる方に

本誌に掲載された著作物を複写したい方は,(社)日本複写権センターと包括複写許諾契約を締結されている企業の従業員以外は,著作権者から複写権等の行使の委託を受けている次の団体から許諾を受けて下さい.

著作物の転載，翻訳のような複写以外の許諾は，直接本会へご連絡ください.

〒107-0052 東京都港区赤坂 9-6-41 乃木坂ビル 学術著作権協会

TEL: 03-3475-5618　FAX: 03-3475-5619　E-mail: naka-atsu@muj.biglobe.ne.jp

アメリカ合衆国における複写については，次に連絡してください.

Copyright Clearance Center, Inc. 222 Rosewood Drive, Danvers, MA. 01923, USA

TEL: 978-750-8400　FAX: 978-750-4744　URL: http://www.copyright.com/

情報知識学会誌 Vol.19, No.2　2009 年 5 月 16 日発行　編集・発行　情報知識学会

頒布価格 3000 円

情報知識学会(JSIK: Japan Society of Information and Knowledge)

会長　根岸　正光

事務局

〒 110-8560 東京都台東区台東１－５－１ 凸版印刷（株）内

TEL: 03(3835)5692　FAX: 03(3837)0368　E-mail: jsik@nifty.com

URL: http://www.jsik.jp

Journal of Japan Society of Information and Knowledge

Contents

Constructing Authority Control File Using Topic Map — Norio TOGIYA,Motomu NAITOU 57

Digital Archives at University Libraries in Japan — Yoshinori SUZUKI,Soichi TOKIZANE 63

Sharing curatorial infomation among museums and developing "Inter-Museum Loan" system — Satoshi TARASHIMA 70

Development of the Proverb Quoting System using Natural Language Processing — Kosuke MAKINO,Yuya KAJIKAWA 74

Sentence Length in Children's Books — Akira YAMAMOTO,Yoko KUSAKA 80

Web API for Biology — Yasumasa SHIGEMOTO,Yoshikazu KUWANA,Yeondae KWON,Hideaki SUGAWARA 86

Development of Bio-metadatabase and its applications — Tomonori SUZUKI,Satoru MIYAZAKI 92

Identification Algorithm of Associated Genes Specific to a Particular Disease — Yeondae KWON,Hideaki SUGAWARA 98

Development of network database system for thermophysical properties data equipped with integrated system — Yuichiro YAMASHITA,Tetsuya BABA 104

Design and Application of Database for Materials-Risk-Index — Hiroshi YOSHIZU,Kohmei HALADA,Mitsutane FUJITA 112

Towards a Quantitative Research Approach to Literary Review: Quantitative Analysis of Book Reviews — Hajime MURAI,Akifumi TOKOSUMI 120

Structual Analysis for an Early Trilogy by Haruki Murakami — Akira KUDO,Hajime MURAI,Akifumi TOKOSUMI 126

Extracting Distinctive Vocabularies and Conceptual Categories from a Literary Text: A Quantitative Analysis of Hoshi Shinichi's Short-Short Stories — Chie SATO,Hajime MURAI,Akifumi TOKOSUMI 132

Conceptual Transitions in Papers of Music Criticism: Extracting Musical Concepts based on Network Centrality — Akihiro KAWASE,Hajime MURAI,Akifumi TOKOSUMI 138

States of mind and stylistic features: The effects of affiliate advertising — Kaori SAITO,Hajime MURAI,Akifumi TOKOSUMI 144

Optimization of automatic assignment of "Kansei" parameters to the books based on the words in online book reviews — Takashi HARADA,Masaaki IKEUCHI 152

A Trend Analysis on the Disparities among Japanese Universities Observed with the Number of Publications and Citation Counts — Masamitsu NEGISHI 158

Extraction of Interdisciplinary Papers by Network Indicators — Yuya KAJIKAWA,Junichiro MORI 170

University-Industry Collaborations in Academic Websites: A Case Study of Seven National Universities — Sumio KAKINUMA,Yuan SUN,Masaki NISHIZAWA,Keizo OYAMA,Masamitsu NEGISHI 174

Interannual Change of Bioinformatics Analysis Techniques by Workflow Extraction from Biological Papers — Jiro ARAKI 179

Co-authorship Prediction based on collaborative filtering — Yukio HORI,Takashi Nakayama,Yoshiro Imai 185

Modeling Knowledge Transfer through Inter-Generational Research Collaboration — Takaki FUKUDA,Katsuhiro UMEMOTO 191

Complete rationality and definitive rationality — Tetsutaro YASUHIRA

Aspects of Information Process — Keisuke TANATSUGU 212

Graphical Expression of SQL Statements using Clamshell Diagram — Takehiko MURAKAWA,Masaru NAKAGAWA 218

Analysis of User Eye Movements during Viewing of Search Engine Results Pages in Web Information Seeking Tasks — Masao TAKAKU,Yuka EGUSA,Hitoshi TERAI,Hitomi SAITO,Makiko MIWA,Noriko KANDO 224

情報知識学会誌 第19巻2号　2009年5月16日発行

編集兼発行人 情報知識学会 〒110-8560 東京都台東区台東1-5-1 凸版印刷㈱内

TEL:03(3835)5692 FAX:03(3837)0368 E-mail:jsik@nifty.com

URL: http://www.jsik.jp/　　(振替：00150-8-706543)

学術刊行物 ISSN0917-1436