T2R2 東京工業大学リサーチリポジトリ
Tokyo Tech Research Repository

論文 / 著書情報
Article / Book Information

論題(和文)	大規模知識資源の体系化と活用基盤構築
Title	
著者(和文)	古井 貞熙
Author	SADAOKI FURUI
出典(和文)	日本音響学会誌, Vol. 63, No. 12, pp. 731-737
Journal/Book name	, Vol. 63, No. 12, pp. 731-737
発行日 / Issue date	2007,

Powered by T2R2 (Tokyo Institute Research Repository)

解説

大規模知識資源の体系化と活用基盤構築

——文理融合で知識資源の時代を開く——*

古井貞熙（東京工業大学大学院情報理工学研究科計算工学専攻）**

43.72.Ne

1. 知識資源の時代に向けて

21 世紀は知識資源あるいはコンテンツの時代といわれており，あらゆる生活面において，大規模知識資源の構築と活用が不可欠になる。これまでの情報化社会の進展を振り返ってみると，1970 年頃に大型コンピュータを多数の利用者が利用する時代（システム中心時代）が始まり，1980 年頃からパーソナルコンピュータ（PC）が作られるようになって，1985 年頃に，大型コンピュータよりも，小さなコンピュータを各個人が持つ時代（PC 中心時代）が到来した。1990 年頃からは，インターネットを中心とするネットワークの重要性が意識されるようになり，コンピュータは単独で使われるものではなく，ネットワークにつながることによって，新たな価値を生み出すようになった。2000 年頃からは，コンピュータそのものよりも，ネットワークをいかに活用するかが，より重要な時代（ネットワーク中心時代）となった。このネットワーク中心時代の次に来るのが，知識資源中心時代である。ネットワークが知識資源を物理的に結びつけるのに対して，知識資源中心時代の課題は，知識資源をいかにして意味・概念のレベルで結びつけ，体系化するかである。

「知識」と関連のある言葉に，「情報」と「データ」があり，これらは 3 階層を成している。データを抽象化したものが情報，それを更に一段高いレベルで抽象化したものが知識で，知識そのものがいろいろな形で抽象化され，体系化される。知識は，いわゆるコンテンツが観測される規則性や，一般的なルールの表現形態であり，それを用いて推論を行い新たな情報を創出したり，情報に対する解釈を与えたり，問題解決をする源となるものである [1]。知識は特定の話題に関する情報が包括的に集積されたもので，情報よりも重い存在である。知識をいかに体系化するかによって，我々が外界から得ることができる情報が，全く変わってくる。情報をどうやって生かすかは，我々個人あるいは社会の中に，知識がどのように体系化されているかによって決まる。知識資源とは，大量の知識を利用できる形で蓄積したもので，メタ知識，すなわち知識に関する知識を含むという意味で，コンテンツよりも広い概念を指す。

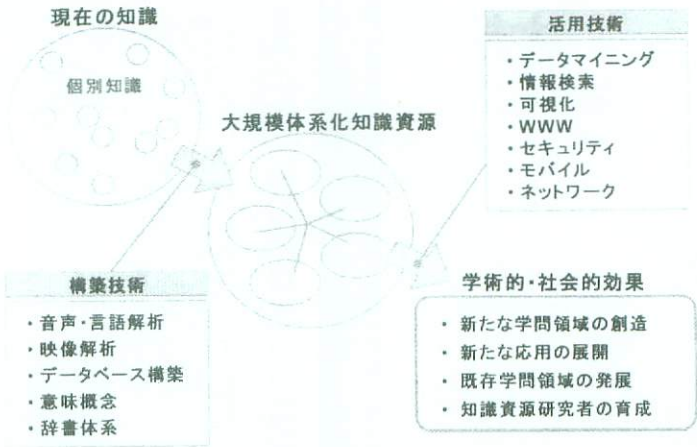
2. 大規模知識資源 COE の研究拠点形成

このような背景から，「大規模知識資源の体系化と活用基盤構築」の COE 研究拠点の形成が計画され，平成 15 年度に開始された。大規模知識資源を構築し活用できるようにするためには，知識自体の集積のほかに，様々な基本技術の創造が必要である。本 COE では，東京工業大学大学院情報理工学研究科，社会理工学研究科，及び学術国際情報センターを中心とする 20 名の事業推進担当者を核に組織を構成し，COE 特任教授（専任及び客員），助手，ポスドク研究者，博士課程学生を加えて，人文社会系・理工系を融合（文理融合）した多様な学際的研究を行っている。具体的な知識資源として，Web，話し言葉，書き言葉，マルチメディアコンテンツ，遠隔教育教材，マルチメディア教材，古典文献，考古学知識などを対象とし，並行して，拠点形成のための基礎となる大規模計算基盤及び大規模情報蓄積・検索・流通基盤の整備を進めている。

本 COE の拠点形成の目的を，図–1 に示す。上で述べた多様な知識資源を対象とし，大規模知識の標準的な体系化と構造化のための枠組みを構築

* Framework for systematization and application of large-scale knowledge resources: Opening the knowledge-resource era by fusing humanities and technology.

** Sadaoki Furui (Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Tokyo, 152–8552) e-mail: furui@cs.titech.ac.jp



図–1 「大規模知識資源の体系化と活用基盤構築」COE の目的

している。多様な知識資源を相互に関連付け，体系化するための意味体系に関する研究と共に，体系化された大規模知識資源の構築を進めている。これにより，誰でも容易に知識資源を構築し，体系的に活用することを可能にする基盤を確立することを目指している。同時に，開かれた拠点として知識資源を蓄積・活用し，新しい知見を創造することで，既存学問のより一層の発展と，境界領域あるいは領域横断型の新しい学問の開拓を目指すと共に，知識資源研究者の育成を目指している。本 COE から得られる新たな知見は，音声・言語学，メディア学，教育学，歴史学，文献学，考古学などの，新たな発展を促すと期待される。本拠点が対象とする知識資源には，古典文献，古典芸術など，我が国の伝統，文化，言語に直接関連するものが多くあり，本拠点での研究により，それらの知識を大量に蓄積すると共に，知識資源の体系化により，従来の個別知識のレベルでは不可能であった新しい事実の発見を目指している。

図–2 「大規模知識資源の体系化と活用基盤構築」COE の研究課題の体系

3. 四つの知識資源グループによる拠点形成

本拠点形成が対象とする，大規模知識資源の体系化と活用基盤構築のための具体的研究課題の相互関連を，図–2 に示す。大規模知識資源の蓄積，計算基盤，及びネットワークのインフラストラクチャを基盤として，オントロジーをはじめとする知識資源の体系化のための基本的な原理やアルゴリズムを研究している。それに基づいて，音声，言語，映像，Web に代表される，素材としての知識資源の構築法を研究し，実際に種々の知識資源を構築する。それらをベースとして，多様なアプリケーションを対象とした，知識資源の活用法について研究し，実際に種々の応用システムを構築する。更に，これらの多様な知識資源を用いた，文理融合による種々の分析を行う。

拠点形成を組織的に進めるため，言語・文献知識資源グループ，Web 知識資源グループ，教育・学習知識資源グループ，音・映像・感覚知識資源グループの，四つのグループを構成して，活動を進めている。各担当者は，必ず一つあるいは関連する複数のグループに属している。以下に，各グループの活動の概要を紹介する。

3.1 言語・文献知識資源グループ

3.1.1 形式オントロジー

オントロジーとは，概念の関係である。概念には通常幾つかの語彙が対応するため，語彙と語彙の関係ととらえる立場もある。また，その関係が持つ意味を，単なる事実のみで与えるだけでなく，その概念間の関係規則を与えることも行われる。例えば，生物情報に関する静的な属性に基づいた情報検索及び，未知の有効な薬剤の効果や副作用の発見において，分子や薬物，生体物質などの相互作用に関する情報を用いることが注目されている。しかし，その情報を表現する言葉の持つ意味は，実際の科学的反応が持つ性質に基づいているにもかかわらず，その詳細を知ることができないことと，その作用と効果の捉え方が一様でないため，曖昧である。一方，共通の言葉が，様々な異なる物質レベルで使われる。この共通性の中には，異なるレベルの現象を統一的に理解するオントロジーが潜んでいると考えられる。本研究では，相互作用の記述に汎用的に用いられる，活性作用，阻害作用の概念を統一的に記述する形式オントロジーを提案した。更に，このオントロジーの推論体系を用いて，世界最大の薬害の存在が証明可能であることを示した。

3.1.2 感性オントロジーと学術知識オントロジー

言語的知識資源の効果的な利用を可能にするために，人間と機械の双方が利用でき，かつ必要に応じて詳細で緻密な記述を提供するオントロジーを，感性関連領域と学術知識領域で構築している。感性オントロジーでは，以下の三つの下位領域のオントロジーを構築している。感情オントロジー：悲哀感情を中心としたオントロジーを構築し，辞書分析，語彙が持つ多義的な意味の定量的解析，口語性の高いテキストデータの分析心理実験により，意味記述の精密性を高めた。意図オントロジー：語用論的意味を分析するための認知的構成素の提案を行い，日本語の遂行動詞（話者の行為を示す動詞）について，分析の有効性を確認した。人格オントロジー：人事管理に用いられる情報を，人間知識資源と考える発想で，人間の性格，能力に係わる概念体系を構築している。学術知識オントロジーでは，下位領域として，認知科学オントロジー，大規模知識資源学オントロジー，及び批評学オントロジーを構築している。

3.1.3 ドメインオントロジーを用いたソフトウェア要求獲得支援ツール

ソフトウェアの開発者が，ユーザの要求分析を行う場合，ソフトウェアが扱う分野についての知識（ドメイン知識）を十分に持っているとは限らないため，これを計算機によって補うことが有用である。本研究では，ドメイン知識としてオントロジーを利用し，その支援ツール及び，テキストマイニング技術を用いた特定のドメインに関するオントロジーの構築を支援するツールを実装した。

3.1.4 単語の関連情報を用いた意味ネットワーク

単語を一つの点，単語と単語の関連を一つの辺に見立てて，言語の世界を巨大な規模のネットワーク（グラフ）の形で表し，単語と単語の関連情報にグラフ理論，特にグラフクラスタリングを適用して，概念クラスターを生成する研究を進めている。古典ギリシャ語による新約聖書の単語意味ネットワークを構築し，マルコ，ルカ，ヨハネの三つの福音書が書かれた起源に関する分析を進めている。

3.1.5 構文木付きコーパス作成支援統合環境

近年，自然言語処理の分野では，大規模な言語資源に基づく統計的手法が研究の中心になっており，特に構文木付きコーパスは，確率的構文解析モデルの学習データや，構文解析システムの評価用テストセットなどに用いられ，重要な言語資源である。しかし，構文木付きコーパスを作成するには，多くの人手と時間を必要とする。このため，その作成支援環境「eBonsai」をベースに，ノウハウを共有する支援を加えた作成支援環境の構築を目指し，システム側が作業過程を統制する手法を提案すると共に，新たなインタフェースを用いた被験者実験を行い，効果を確認した。

3.1.6 比喩理解と比喩生成

人は文学的表現に限らず，科学的記述や日常生活での会話においても，様々な種類の比喩を用いている。直喩に限定しても，「雪のような肌」「バレリーナは蝶のようだ」などの表現を，人はごく

日常的に理解し，生成する。そのため，比喩理解・生成を表現するモデルは，人が用いる概念を十分に網羅し，心理的妥当性を確保することが必要である。本研究では，言語統計解析を用いて作成された確率的知識構造に基づいたモデルの構築と，その妥当性に関する心理学実験を行っている。具体的には，毎日新聞 10 年分から抽出した形容詞と名詞の係り受け頻度データを対象として確率的潜在意味分析を行い，ニューラルネットワークを用いてモデル化を行った。実験の結果，モデルの妥当性が実証されている。

3.2 Web 知識資源グループ

3.2.1 ニュースディレクトリの構築と検索

地球上の 20 箇所以上のニュースサイトの 4 か国語の自然言語で記述されたニュース記事の索引情報を，毎日自動収集し，従来の Google ニュースなどのキーワード検索をベースとするニュース検索では検索が困難だった場合のニュース検索も一部可能になる，ニュースディレクトリシステムの構成法と検索法を提案した。従来のキーワード検索をベースとする検索では，母国語だが呼び名が分からない場合や，知らない外国語で検索用語が分からない場合などの検索ができなかった。本システムでは，検索で数を絞り込んで，ユーザの知らない外国語のニュース記事の場合でも，自動翻訳無料サービスなどを利用し，記事の内容を読むことを可能としている。

3.2.2 インターネット QA サイトにおけるリンク予測

近年，インターネット上には，人間同士がコミュニケーションをとる場が多く存在し，QA（質問応答）サイトもその一つである。QA サイトは，ある人が投稿した質問に対して，他者が回答を寄せるシステムで，そこでの回答者のつながりは，ソーシャルネットワークと捉えることができる。このようなソーシャルネットワークにおいて，将来できるであろう新たな人間関係（リンク）を予測することは，コミュニケーションの促進につながり，重要である。本研究では，ネットワークのノード間の類似度を測る従来手法を改良することによって，新たにリンクの重みを考慮したリンク予測手法を提案した。Yahoo!知恵袋のデータから構成したネットワークに適用した結果，従来法よりも精度が向上することが確認された。

3.2.3 統計的方法による QA システム

オープンドメイン，ファクトイド質問をタスクとした，データ駆動，統計的アプローチによる QA システムを構築し，英語，日本語をはじめとする多言語を対象としたシステムで，TREC などの国際ワークショップに参加して，高い評価を得ている。

3.3 教育・学習知識資源グループ

3.3.1 日本語読解学習支援システム「あすなろ」と日本語作文支援システム「なつめ」

日本語学習者のための，多言語によって語の意味や文の構造が提示できる，学習支援システムを構築した。Web 上で公開中（http://hinoki.ryu.titech.ac.jp）で，学内外で利用されている。学習者が入力した日本語の文章に対し，文章中の単語の訳と文ごとの構文構造を，日本語，英語，マレー語，インドネシア語のほかに，中国語，タイ語等の特殊な文字を含めた多言語表示ができる。

更に作文支援のため，文学作品，論文などから，学習者の日本語習得レベルに応じて，単語難易度，文法難易度に関して適切な例文を抽出し，表示することができるシステムを構築した。名詞，格助詞，述語の共起情報を検索して，例文を得ることもできる。

3.3.2 論文データベース統合システム「PRESRI」

特定分野の研究動向を知るためには，その分野の論文を網羅的に収集する必要がある。このような文献調査を行うのに，しばしば論文データベースが用いられるが，それらが分散して存在していると，個々に検索するのは非効率的である。そこで，研究論文の参考論文情報を有効利用して，論文間の関係や研究の流れを抽出する「PRESRI (Paper Retrieval System using Reference Information)」システムを構築した。多様なソースからの論文について，論文間の参照・被参照関係をもとに，ある分野の論文集合を集めて表示すると共に，各論文のアブストラクトが読め，ある論文を参照する他の論文中で，その論文に関して記述している部分（参照箇所）を読むことができる。参照の目的に関して，手がかり語を用いて，問題点指摘型（他の論文の理論や手法等の問題点を指摘する），論説根拠型（ある理論を提案する場合や仮定をする場合にその根拠となる論文）及びその他に，自動的に分類している。

3.3.3 教育コンテンツ統合システム「UPRISE」

インターネットで情報発信を行い，かつ大量の非定型データやマルチメディアデータを扱うアプリケーションの一つとして，e-ラーニングが注目されている。本研究では，教育コンテンツの柔軟な統合機能と高度な検索機能の実現を目指し，大学での講義のビデオ画像を中心とする多様な情報を，実際に使われている状況の知識を有効利用して統合的に蓄積することにより，キーワード検索可能にする「UPRISE (Unified Presentation Contents Retrieved by Intelligent Search Engine)」システムを構築した。

UPRISEでは，メタデータによるコンテンツの統合のために，動画ストリームをシーンの連続であると抽象化し，各シーンとそこで使用されたスライド資料とを対応付けることでそれらを統合化する。また，各シーンに対して，対応する資料の文字・構造情報，シーンの長さ情報，音声情報，レーザポインタのポインティング情報から，検索用インデクスを作成し，高度な検索を可能としている。スライドの切り替えタイミングによってシーン分割を行うため，同じスライドが繰り返し使われても，違うシーンとして区別している。

講義音声を認識して音声情報を得るための，音素モデル及び言語モデルは，日本語話し言葉コーパス（CSJ）中の講演音声を用いて作成し，単語辞書には，各講義のスライドから抽出した名詞を追加登録した。音素モデルを，MLLR法により話者に教師なし適応すると共に，講義で使用されたスライド資料を用いて言語モデルを動的に適応することにより，キーワードの抽出精度の向上を実現している。

3.4 音・映像・感覚知識資源グループ

3.4.1 放送番組の検索・要約

テレビ放送などに使用されるマルチメディアコンテンツは，日々増大の一歩を辿っており，その蓄積量は膨大なものとなっている。その有効活用のための要約・検索手段の開発を目的として，NHK放送技術研究所と共同で，野球放送ビデオからのシーン（イベント）の自動抽出の研究を進めている。野球放送のデータ構造の最小単位はフレームで，1枚の静止画像であり，一つの固定カメラで撮影された多数のフレームからショットが構成される。更に，ショットのシーケンスからシーンが構成される。ショットの遷移情報はシーンの特徴を表し，シーンの認識において重要な情報となる。ゲームの内容を理解するために重要となるシーンは，ハイライトと呼ばれる。

連続音声認識とのアナロジーにおいて，ショットを音素に，シーンを単語に対応させ，シーンの遷移を統計的言語モデルで与えることにより，統計的音声認識の枠組みでシーン認識を行った。PCAによる主成分特徴など，種々の画像特徴を用いた実験の結果，この方法の有効性が確認された。更に，音響特徴をマルチストリームHMMで組み合わせた結果，ハイライトの認識に有効であることが示された。

3.4.2 話し言葉音声の分析と要約

話し言葉音声の全体像を適切に把握するために，書かれたテキストを読み上げた音声と，通常の話し言葉音声の音響的・言語的違いの分析を進めている。これまでに，音響的特徴としては，話し言葉音声特有の「発声の怠け」によって生じる音素ケプストラム空間の縮小，言語的特徴としては，言語モデルにおけるテストセットパープレキシティの増加が，話し言葉音声の認識性能の低下の大きな要因であることが確認された。

話し言葉音声には，言い直し，言い誤り，間投詞など，冗長な表現が多く含まれている他，それを音声認識した結果には，認識誤りが避けられない。音声ドキュメントの検索などにおいては，音声をそのまま録音したり書き起こしたりするだけでなく，要約したテキストや，要約音声で提示することが求められることが多い。これまでに，音声認識結果に対して，重要文抽出と単語抽出の組み合わせによって，自動的に冗長な部分や認識誤りを除き，重要な部分だけを抽出することによって音声を要約する方法を提案し，その有効性を示すと共に，要約性能を定量的に評価する方法を提案した。

なお，日本語話し言葉コーパス（CSJ）は以前からXML化されているが，種々のユーザの利便性を考慮してRDB化を進め，本COEで構築した大規模知識資源蓄積システムに蓄積して，外部の研究者などが検索・利用できるようにした。

3.4.3 古典文学と物語映画の時空間分析

日本人は8世紀以来，多くの古典文学を文化的

財産として継承してきている。文学は人間の知識と知能の基礎となる言語そのものであり，日本人の言語生活の歴史と変遷をその中に窺い知ることができる。本研究では，文理融合により，古典文学に統計的あるいは数理的手法を適用することによって，音韻，音節，語，韻律的特徴などの日本語の音声的諸特徴について，新たな知見を得ることを試みている。

韻律では，平家物語の時代に七五調と呼ばれる語りの形式が確立したが，七五調は，俳句や和歌など日本の伝統的詩歌の形式ともなっており，現代でも日本人が好む代表的リズム形式である。七五調は，8拍4ビートの当時的リズムが基本となっており，時間軸上では8拍長と5拍長のリズムであるため，両者の比は黄金分割比となる。七音と五音からなる日本の伝統的詩歌の俳句と和歌についても，その基本的分割点で2分された区分間の比は，同じように黄金分割比となることが見出された。

等時的基本リズムの上に構成されるこのような不等長リズムは，他の芸術分野においても好まれていると考えられる。そこで，日本が生んだ有名な映画監督の一人である小津安二郎の映画について，その時間的・空間的特性を調べた。その結果，ショット長や画面の構成原理に関して，対称的なバランス構造に，黄金分割比に従う非対称的構造を重ね合わせた調和的空間を実現していることが見出された。

3.4.4 考古学情報抽出支援

考古学は，研究対象に関する膨大かつ多分野にわたる情報を収集整理し，それらから過去の様相を復元する研究である。その研究材料となる情報は，遺跡発掘調査だけでも，その数は日本国内で年間数万件にのぼる膨大なものである。この情報量は，既に個人の能力で処理できるものではなく，情報の欠落を生み易い状況にある。そのため，膨大かつ多様な考古学情報を有し，考古学研究者が実際に必要な情報を得ることができる，データベースシステムの構築が求められている。本研究では，そのようなニーズに答えるため，「ARCADIA」システムを構築している。考古学情報の近接性に着目し，研究者が必要としていると考えられる互いに関連した情報群を抽出する仕組みになっている。

4. 大規模知識資源の蓄積及び計算基盤の構築

4.1 大規模知識資源蓄積システム

本 COE では人文社会学系・理工学系の研究者が連携して知識資源を構築しており，多種多様な大量の知識素材を蓄積し，それらを整理しながら利用して研究を進めている。幸い近年の情報技術の発達により，テキスト，図表，画像，音声，動画など知識資源の元となる多くの素材が電子化され，情報環境を利用してそれらを蓄積，検索することが可能となってきている。本 COE では，そのような知識資源構築のための研究基盤として，大規模知識資源蓄積システム (KS；Knowledge Store) を開発した。KS では，様々な利用形態や多様な素材の蓄積，更に格納手法に関する試行錯誤等を可能とするため，柔軟性・拡張性に重点を置き，共通する基本的な蓄積・検索機能を提供し，利用分野毎に外部システムを用意する方針を採っている。その内部では，各素材を統一的に扱い高度な検索を可能とするため，素材に関する情報であるメタデータを柔軟に定義することを可能としている。また，外部システムとの間はネットワーク環境での利用を想定し，標準化されている Web サービスインタフェースを提供している。利用者は外部システムを通し，あるいは直接 Web インタフェースで KS を利用することができる。

4.2 大規模計算基盤システム

本 COE では，様々な形式の自然言語，画像，音声に関する信号処理，知識処理，及びデータマイニングを行う。また，従来型の数値処理や，Web に対する自動情報収集なども行われる。セマンティック Web のデータマイニングでは，数ギガバイトから，最大では数テラバイトのデータ量を一度に扱うことが必要となり，また，自然言語マイニングや音声認識理解の計算量は莫大となる。学内に分散した，本 COE に関連する 20 の研究室が，有効に知識処理の対象となるテキストや音声データを作成・共有し，それに対して負荷の高い処理を行うのは，従来の研究室のパソコンなどでは不可能である。そこで，データ処理中心のグリッドであるデータグリッドのアーキテクチャを，大規模知識処理をアプリケーションターゲットとして構築し，それを実際の大規模計算基盤として実現することによって，研究遂行の礎とした。COE にお

ける研究活動での実運用を通じて，高信頼性・冗長性・可容性・高いアクセシビリティの実現のための技術的要件の洗い出しと，学内キャンパスグリッドを中心とした他の情報インフラとの連携を実現している。

5. 夢の実現へ

本COEでは，これまでに述べたように，文理融合という旗印のもとに，人文系スタッフと情報系スタッフの融合によって，大規模知識資源を構築し，それを活用流通するための，新しいアルゴリズムや体系化の研究と教育を進めている。この拠点形成のためには，国内のみならず国際的先進研究機関や大学と協調することが必要で，欧米を中心とする多数の機関との協力関係を進めている。「Symposium on Large-scale Knowledge Resources」[2, 3] を毎年，国際会議あるいは国内会議として開催し，国内外からの招待講演者やCOEメンバーの講演を中心に，活発な討論を行っている。スイスを中心とする欧州のプロジェクトとの，共同シンポジウムも開催している。また，活動の促進のため，20名余りの博士課程学生をRAとして雇用し，関係する博士後期課程の中に，特別コースを創設すると共に，博士課程学生の自主的な活動の基盤としての「博士フォーラム」を進めている。多数の学生が，成果の発表，討論，共同研究などのために，海外に派遣されている。本COEがねらう大規模知識資源の体系化と活用基盤構築は，次世代Webとも関連して，すでに世界的なレベルでのホットな研究課題となりつつある。

文　献

[1] 西尾章治郎，大田友一，横田一正，西田豊明，佐藤哲司，“情報の共有と統合，”岩波講座マルチメディア情報学，Vol. 7（岩波書店，東京，1999）.
[2] *Proc. Int. Symp. Large-scale Knowledge Resources* (*LKR 2006*), Tokyo (2006).
[3] *Proc. Symp. Large-scale Knowledge Resources* (*LKR 2007*), Tokyo (2007).