連載「情報学を創る」
―科研プロジェクトがめざしたもの

コンテンツの生産・活用に関する研究

― 科研「情報学」プロジェクトの
コンテンツ研究を振り返って ―

安達　淳（国立情報学研究所・
コンテンツ科学研究系）
喜連川優（東京大学生産技術研究所）
中川裕志（東京大学情報基盤センター）

文部科学省科学研究費補助金の特定領域研究「IT の深化の基盤を拓く情報学研究」（略称：「情報学」，領域代表者：安西祐一郎）の中で設定されたコンテンツにかかわる研究項目，「A02　コンテンツの生産・活用に関する研究」で行われた研究の概要と，その中の研究例をいくつか紹介する．

コンテンツ研究項目の設定した研究領域

20 世紀末に出現し爆発的に発展した Web 文書を代表とするデジタル情報は，現在，きわめて多くの利用者が世界的規模で日常的に活用するコンテンツとなっており，情報の一様性に欠くといった問題はあるものの，今後の社会基盤を形成していくものであると同時に重要な知識基盤の一翼を担うものと捉えられている．

しかし，情報量の大規模化，情報の言語的・メディア的・文化的多様化とそれに呼応する情報ニーズと利用行動の多様化，情報形式の断片化や生産・消費の高速化，情報の利用者が同時に生産者でもあるという相互性の進展などの実態に即した，ディジタルコンテンツの生産・活用・処理については，技術ばかりでなくその概念自身も十分には明確にされていない．

2001 年 10 月から 2006 年 3 月まで実施された特定領域研究「情報学」の 6 つの柱の中の 2 つ目の研究項目は，A02「コンテンツの生産・活用に関する研究」と称され，上記のような問題意識のもとで，情報処理技術の基本である情報の管理と利用に関する研究と諸技術，すなわちデータベース管理システム，マルチメディア情報処理，高速大量データ処理，分散データベース，データマイニングなどのデータ工学分野，自然言語処理や情報検索などの情報活用技術，地理情報や画像情報などのメディア工学技術，情報利用や流通モデルといった社会情報研究等の分野の研究を推進してきた．

この研究項目ではさまざまな研究を**図 -1** に示すように捉えた．Web という重要な情報資源の活用を巡り，(1) 種々の不均質性を持つコンテンツの特質，(2) 形式的操作を主体に進んできたデータ工学関係の研究と，内容に即したかたちで進んできた自然言語処理や情報検索といった研究などを総合的に行うこと，(3) コンテンツを活用していく観点から，Web 情報の社会的あり方も考慮しつつ種々の活用を目指すこと，という視点を持ち，

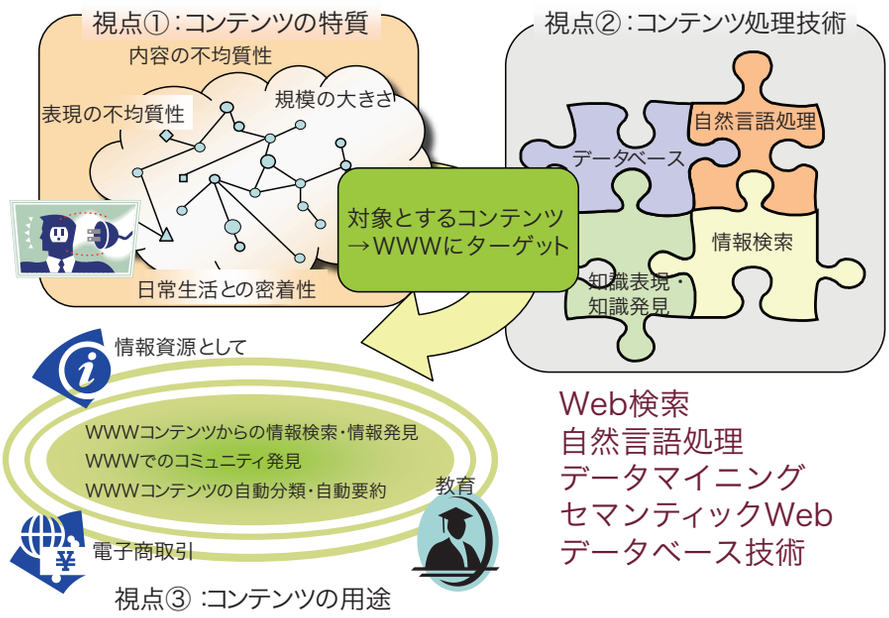

図 -1
科研におけるコンテンツ研究の構成

この研究によりシステムとニーズ，コンテンツに関する技術と理論の接点で新たな情報学研究の前線をかたち作ることを目指した．また，Web コンテンツだけを狭く捉えるのではなく，近年社会的にもきわめて重要になってきている XML などの半構造データや地図情報なども視野に入れて，情報を統合的に活用する方法を探る研究もカバーしている．

以上のような大枠の目標のもと，5 年に渡る研究期間全体の前半 3 年においては，公募研究の持つ多様性や独創性を伸ばしつつ，研究班のインタラクションを活性化することを試みた．そして，後半 2 年は成果を収斂させ分かりやすい成果としてアピールしていくことを最重要課題と設定した．

研究経過

特定領域は当初からの計画に含まれている計画研究というグループとテーマに呼応して応募し審査会で選定される公募研究から構成される．A02 は，当初は計画研究 2 班と公募研究 21 班で開始し，2 年目は公募研究 19 班，3 年目は公募研究 17 班，そして最終の 2 年間は公募研究 20 班で構成されている．このように毎年公募研究については何件かの入れ替えがあり，どのような公募研究が採択されるかは，領域運営側ではなく審査会側によって決められる．各研究班のテーマはコンテンツにかかわる研究要素を複合的に含むものであり，排他的にいずれかに整理することは難しいが，あえて，重複を厭わず，各研究班の研究を整理すると大きく 3 つに分けられる．

【I】コンテンツを支える形式的・構造的特質を中心とする研究

このグループは，主としてデータベース技術と Web マイニング技術とに大きく分けることができる．前者では柔軟なデータ操作のための XML 構造の部品化，応用として位置情報や教育情報を扱うものへの成果が上がっている．Web マイニングでは，Web 情報空間の構造に適したデータマイニングや処理アーキテクチャの研究が進められており，基礎技術としてリンクと半構造データを対象としたマイニングに関する成果が上がっている．

【II】コンテンツの内容的特質を反映した研究

内容的特質を中心に考慮した研究としては，Web 検索・活用テストベッドの構築から不均質多メディア・多言語処理を考慮した総合的な研究，Web 検索研究，言語メディア処理，セマンティック Web をはじめとする知識やオントロジー，ユーザコミュニティを扱う研究が多く，それぞれに，基本的な技術と新たな方式の提案が進められた．社会的な情報利用や流通についての検討など，ここに分類される研究では，技術と内容との関係にとどまらず，情報利用形態や流通の問題までカバーしているものもある．

【III】具体的なニーズや用途を想定した研究

Web コンテンツの形式的・内容的処理を巡る研究を進めるにあたって，ときに具体的な用途が処理に強くかかわってくることがある．たとえば，空間コンテンツや教育コンテンツ，設計・製造知識，そして言語コンテンツの中でも活用状況を特化した事典的コンテンツの処理などに関するものであり，上記の 2 つにも重複分類され得るものである．

多用なテーマの集まる研究グループを取りまとめるのは至難であり，常に発散しがちであった．しかし，

- コンテンツの特質を明確にし，活用のための情報処理技術を確立する
- 具体的なアプリケーションの開発を通して技術を検証し，課題を明確にする

という方針を中心に研究を構成し，さらに進んで今後の情報活用システムのあり方を描き出すことも積極的に考えて研究を進めてきた．

研究成果

20 以上にも及ぶ研究班の成果は，数多くの学術論文や特許になっており，それらはすべて Web で公開されているので，文献 1)，2）を参照願いたい．また，電子情報通信学会の英文誌の 2004 年 2 月号，人工知能学会誌の 2004 年 12 月号で本研究項目を中心とした特集号を企画している．また，データ工学分野での最高峰の国際会議である IEEE International Conference on Data Engineering 2005（ICDE 2005）を東京で開催した際には，本研究項目を核に併設のかたちで国際ワークショップ International Workshop on Challenges in Web Information Retrieval and Integration（WIRI）を開催した．18 カ国から 47 本の論文投稿があり，そのうち 11 本がフルペーパー，20 本がショートペーパーとして採択され，そのうち A02 からの採択は，12 論文であった．この WIRI は好評で，引き続き 2006 年の ICDE においても開催された．

A02 の研究成果の別種のエビデンスとして，いくつかの大型の研究が，本特定領域研究とは別の競争的研究資金を得て発足したことが挙げられる．自由な発想や独創性を活かす科研費によって開始された研究の持つ先駆的成果が社会的に認められ，それが別の経費により，より大規模にあるいは実用化を目指して研究開発が進められるということを示すものであり，科研費の持つ特性がよく発揮された研究進化の好例といえる．2003 年度からの「e-Society 基盤ソフトウェアの総合開発」，2004 年度

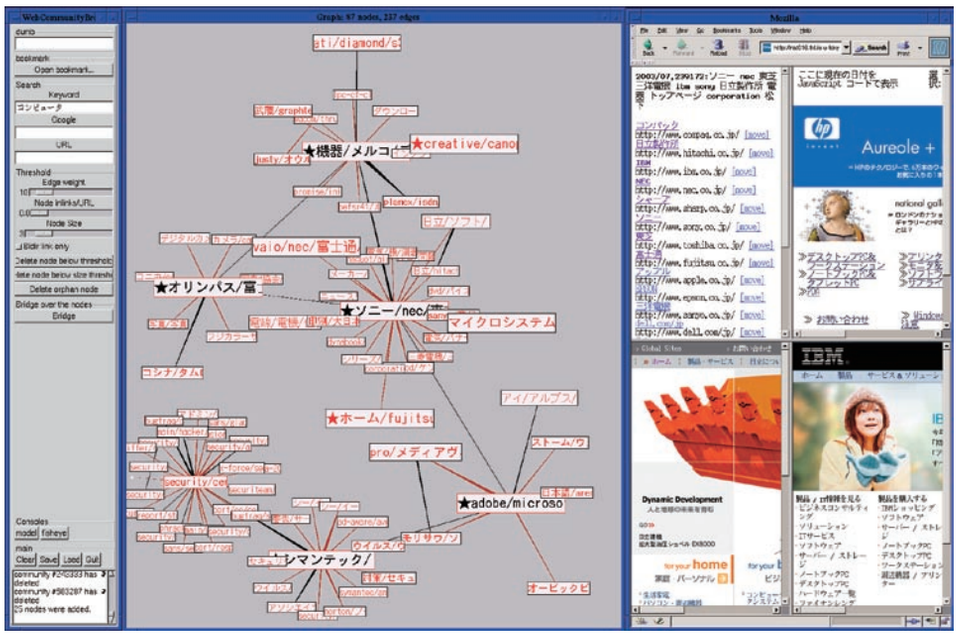

図 -2　Web コミュニティブラウザ

からの「知的資産の電子的な保存・活用を支援するソフトウェア技術基盤の構築」などによって採択された研究がその例である．

本稿では，以下に 2 つの研究例を挙げて，A02 研究項目の典型的な研究活動として紹介する．もちろんこれ以外に興味深い研究例は多いが，それらについては Web サイトを参照していただきたい．第 1 は，大規模な Web コンテンツに対する包括的なチャレンジの好例である．第 2 は，自然言語処理とコンテンツ研究の融合例として紹介するものである．

Web ウェアハウスとコミュニティ抽出

資源の乏しい我が国において，Web 情報は有効活用すべき貴重な資源と見なすことができるが，現状の利用形態は全文検索に基づくサーチエンジンに強く依存し続けている．従来にはない新しい Web の活用法を模索すべく，Web コンテンツを柔軟に操作可能とする強力なプラットフォームを構築することを目標とした研究が行われた．ポイントは，Web ウェアハウスに適合した大規模処理システムアーキテクチャの構築，収集された大容量 Web ページの高度検索処理に適したデータ管理技法，Web ウェアハウスの高次元利用を可能とするログ解析技法，および Web コミュニティの新規出現抽出手法などである [3), 4)]．

【Webウェアハウスに適合した大規模システムアーキテクチャ】

PC クラスタは数多くのコンピュータを高速の LAN で接続したシステムで，そのスケーラビリティやコストパフォーマンスの高さから，次世代大規模並列処理システムとして着目されてきた．一方，Web ウェアハウスは，Web から収集したきわめて大容量のデータを効率よく管理する必要があり，そのプラットフォームとしての整合性が求められる．すなわち，超大規模データを処理する高い計算能力，指数的に増大する傾向に対するスケーラブルな拡張性，超広帯域入出力アクセス性能などが重要な要件となる．

PC クラスタに SAN で結合したストレージを組み合わせることにより，Web ウェアハウスに適合したシステムアーキテクチャを提案し，ノード間転送性能，並列マイニング処理性能等の基本性能を確認した．

【大容量 Web ページの高度検索処理に適したデータ管理技法】

膨大な Web ページは，そのままのかたちではユーザの欲する情報を検索することは非常に困難である．そこで，Web ページのハイパーリンク情報を基にページ間の関連を抽出するアプローチを採用した．Kleinberg によるハイパーリンク解析手法である HITS を拡張することで，Web 上のコミュニティを自動的に抽出し，関連するコミュニティを結んだ相関図（コミュニティチャート）を作成する手法を独自に開発し，収集した全日本 Web ページのスナップショットからコミュニティチャートを作成した．また，抽出されたコミュニティチャートを可視化し，ユーザによる閲覧・探索を支援するツールを構築し，Web コミュニティの検索，解析を容易に行う手段を提供した（**図 -2**）．

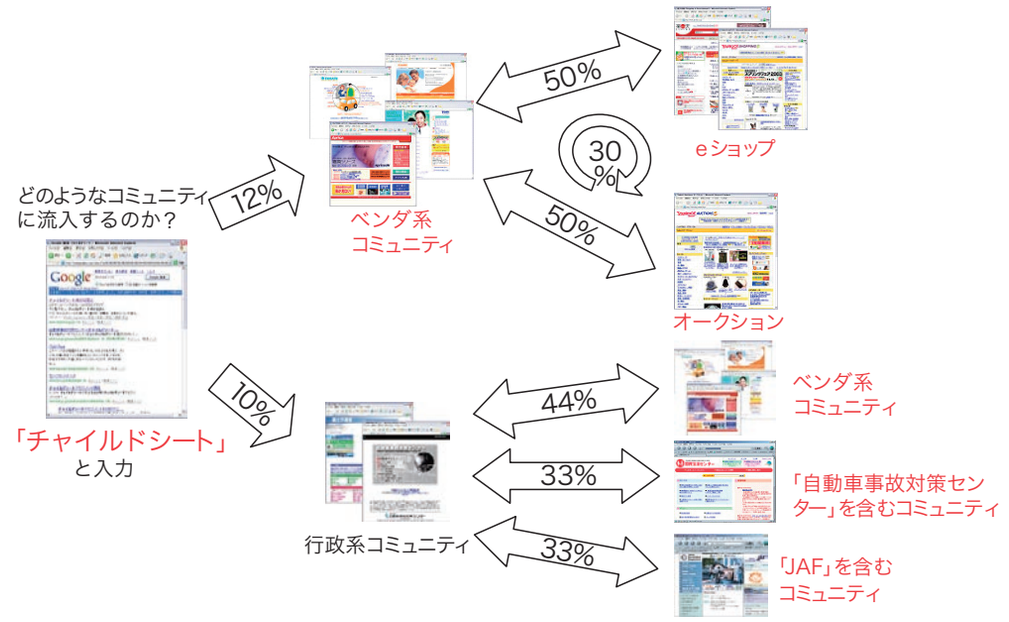

図 -3 「チャイルドシート」を検索したユーザの挙動

【Webウェアハウスの高次利用を支援するログ解析技法】

膨大なサイバー空間におけるユーザ挙動の解析は，個々のサイトのユーザビリティ向上，使用されている検索語からのトレンド抽出，さまざまな挙動の類型からの推薦など，さまざまな応用を持つ重要な研究分野である．本研究では，インターネット視聴率測定のための大域的な Web アクセスログを解析し，新しい大域ログ解析システムの提案を行った．

大域ログにおいては，多数のユーザが膨大なサイバー空間をアクセスするため，URL 単位の解析では大局的なユーザの挙動を捉えることは難しいことから，Web コミュニティ抽出手法を統合したログ解析手法を開発した．コミュニティチャートを用いることで，ユーザの挙動を，URL 単位ではなく，より抽象的なコミュニティ単位で捉えることが可能となった．この大域ログ解析により，ユーザがコミュニティを訪れる目的（検索語）や競合するページ間での目的の違いなど，従来にはない大域的な動きを把握可能とした．

抽出されたユーザ挙動の例を**図 -3** に示す．これは，「チャイルドシート」と検索エンジンに入力した後にユーザが閲覧したコミュニティの解析を行った結果である．チャイルドシートの使用期間は短いためオークションなどで中古品を探すユーザが多く，同時にチャイルドシートベンダとショッピングサイトで性能と販売価格の調査を行う傾向がある．一方，行政関連のコミュニティを訪れるユーザはベンダや日本自動車連盟（JAF）などを含むコミュニティを訪れることから，チャイルドシートの安全性などの調査が目的だと推測できる．

【Web コミュニティの新規出現抽出手法】

Web は敏感に社会動向を反映しており，新しい Web コミュニティの出現を捉えることは，実社会のトレンドを把握する上で重要な課題である．本研究では複数年分の Web スナップショットから抽出された Web コミュニティの出現過程を観測することにより，コミュニティの成長パターンを類型化した．

コミュニティの抽出手法として，HITS を拡張した手法，可変辺容量を利用した最大流アルゴリズムによる Max-Flow 手法の 2 つを用いた結果，Max-Flow 手法で抽出した発生間もない疎なコミュニティが，HITS で抽出できる密なコミュニティに成長する様子を確認できた．またグラフ構造の特徴的な変化を類型化し，コミュニティの成長パターンを明らかにした．

図 -4 は，2001 年から 2002 年の間に出現したアカペラに関するコミュニティの成長を示したものであり，2001 年，2002 年当時のグラフ構造が並べて表示されている．テレビ番組の影響によりアカペラが流行するに従い，多くの著名なページを指す巨大ハブページが増加した結果，大きなコミュニティが形成されていく様子が見て取れる．大規模な調査の結果，こういった巨大ハブの増加による成長パターンが Web 全体で過半数を占めることを確認できた．

Web からの多言語用例検索システム Kiwi

「テキスト」という文字列を Google で検索してみると，瞬時に 355 万件の検索結果があるという表示が得られ

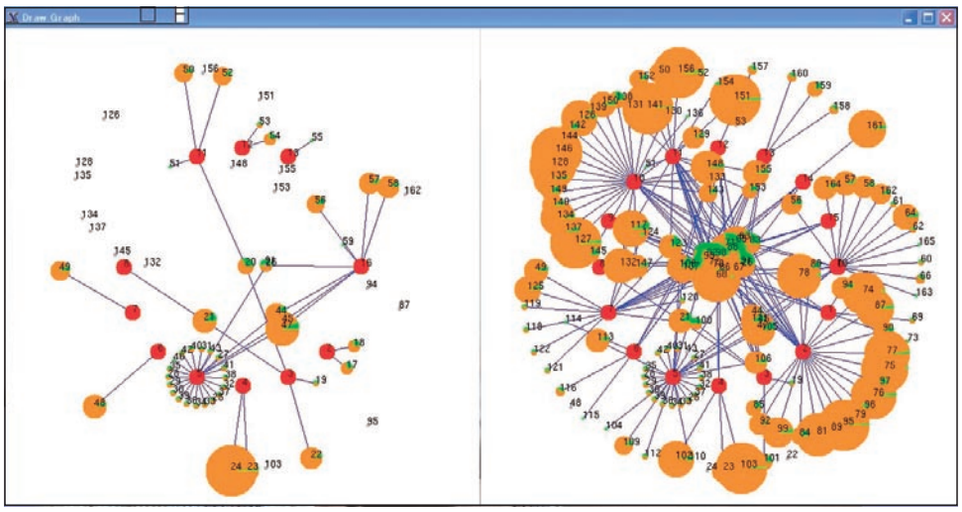

図 -4　コミュニティ出現の実例（2001年〜2002年のアカペラブーム）

赤ノードはコミュニティ内のオーソリティページ，オレンジのノードはオーソリティを指すハブページを表し大きさはコミュニティ外部へのリンクの割合を示す．緑ノブ分は，コミュニティ内部，すなわちオーソリティへのリンク数の割合を示す．オレンジの大きなノードは，リンク数が多いがコミュニティ内部へのリンクが少ないハブページであり，2001 年 にはリンク数の少ないハブにより構成されていたコミュニティが，2002 年には巨大ハブに成長していることがうかがえる．

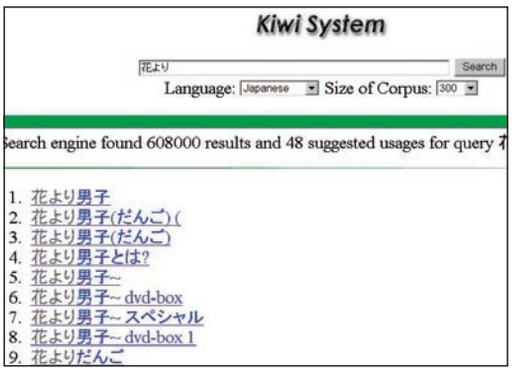

図 -5　Kiwi の検索例：「花より」に後接する表現

る．いくらなんでも 355 万の Web ページを読むことはできない．簡明に検索結果の全容やら，傾向やらが分かる賢い方法を実現したのが Keyword In Web Intelligence, 略称 Kiwi である．

Kiwi では，検索の対象が Web 全体だから，「花より」というフレーズで質問し，「花より」に後接する表現を求めると，**図 -5** のようになる．実際の Web 上のデータでは言い古された「花よりだんご」ではなくて，「男子」という表記を巡る出現が上位である．これは世相を映していて面白い．

【Kiwi の仕掛け】

Kiwi は Web 検索エンジンの検索結果として得られる数行の簡略な文字列 snippet を統計処理して上述の目的を達成しようとするシステムであり，**図 -6** のような情

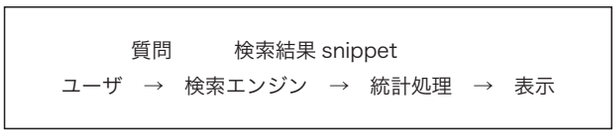

図 -6　Kiwi の情報の流れ

報の流れで処理が進む．検索結果の snippet が得られると，次は役に立つ部分を取り出して表示するための統計処理を行う．Kiwi 独自のアイディアは，この統計処理の部分にある．

まず適当な長さの文字列にして切り出す作業をする．適当な長さだからといって，10 文字とか 20 文字のように固定した長さにすると，重要な表現あるいは単語が途中で切れかねない．また，同じ表現が繰り返し表示されると読むための時間がかかる．そのため，統計的に有意な頻出する表現を取り出して表示したい．要するに，上記の「適当な長さの文字列」とは，意味的にまとまりがよく，頻出する文字列ということになる．

仮に「犬も」という質問に対して，以下の a）〜 d）の 1 文字目が「歩」で始まる 4 本の文字列が検索結果の snippet 群に含まれていたとしよう．また，1 文字目には「歩」のほかに 6 種類の文字から始まる文字列があったとする．

a）歩けば棒にあたる
b）歩けばなんとやら
c）歩くとなんだっけ
d）歩くと飼い主も歩く

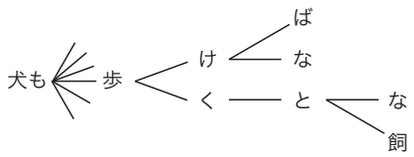

図 -7　Trie 構造の例

すると，**図 -7** のような Trie 構造でこの snippet を表現できる．Trie を見ると，後接する文字種類数が増加に転ずるところまでが意味的にまとまりがよいことが見て取れる．よって，後接する文字種類数が増加に転じたところまでの文字列を切り出す．図 -7 の例だと，「歩くと」が切り出される．

このような方法でまとまった意味の文字列が切り出せるが，snippet の数が多いので，多数の文字列が切り出される．それらを無作為に表示するだけでは，知的でない．良いものから順に表示したいところである．そこで，よく使われるものは重要であること，また長い表現のほうが利用者には理解しやすい，という 2 点を考慮し，実験的に調べた結果，候補文字列（ここでは S と書くことにする）を次の式 K(S) の値の大きい順に表示することにした．ここで，結果の snippet 群における出現頻度を freq(S), 文字列 S の長さ（文字数）を length(S) としている．

$$K(S)=freq(S) \times \log(length(S)) \qquad (1)$$

これまで説明してきたのは，質問文字列に後続する文字列を検索して統計処理して表示するものであった（前方一致検索）．しかし，質問文字列の前方に接続する文字列も検索したい（後方一致検索）．後方一致検索は，ここまで説明してきた方法を，前後逆転して適用すればよい．統語構造を利用しているわけではないので，文字列の構造，すなわち文字の連鎖の仕方の統計的性質は前後反転しても，かなりの程度に成り立つから，前後反転しての適用でもうまく動作する．Kiwi では，これらの検索は，ワイルドカード「*」を使って，たとえば，前方一致検索は abc*，後方一致検索は *xyz のように指定できる．

【Kiwi の使い方と評価】

Kiwi の用例文字列切り出しは，言語に依存しないので，基本的にはいかなる言語にも対応する．応用例として Kiwi は英語で論文を書くときに，自信のない用例の確認や適切な用例表現を効率よく探し出すためのツールとして使える．

「動詞 it seriously」という表現においてよく用いられる動詞を調べるために「* it seriously」という検索を行った結果を**図 -8** に示す．面白いことに上位は take という動詞で独占される．つまり，少なくとも Web 上では，take it seriously という表現はほとんど熟語といえるほどに安定して使用されていることが分かる．

1. take it seriously
2. don't take it seriously
3. taking it seriously
4. to take it seriously
5. if you don't take it seriously
6. please don't take it seriously
7. ever take it seriously
8. never take it seriously
9. took it seriously
10. takes it seriously
11. not take it seriously
12. we take it seriously

図 -8　* it seriously の検索結果　上位 12 位

文献 5）では，現状の Kiwi を “More than search engine, Less than QA” と位置付けている．Kiwi は単語やフレーズを質問として与えて検索する Google などの検索エンジンよりは，検索された内容そのもののテキストとしての特徴をうまく表示してくれる．また，Web 全体を対象データとする検索エンジンの snippet を用いているから，図 -5 に例を示したように，結果は Web の現状を反映したものになっている．だから，“More than search engine” と喧伝しても恥ずかしくはない．

さて，情報検索や自然言語処理の分野でさかんに研究されている質問応答（Question Answering 略して QA）という技術がある．QA は，テキストコーパス中における知識を問うものである．たとえば，「日本で二番目に高い山は？」（百科事典的知識）とか「フランスの歴代大統領は？」（歴史的事柄），さらには「最近，社長が交代した企業名と新旧社長の名前を知りたい」（現代社会における出来事）というような質問の答えをテキストコーパスから検索する技術である．QA をテキストコーパスではなく，Web に対して行うことを WebQA というが，なかなか実現の困難な技術である．Kiwi でこのような QA の問いに答えられるかというのは興味ある問題である．では，実際に上記の質問を Kiwi にしてみたのが**図 -9** である．6 位に正解が現れている．検索エンジンではできないことをやってのけるが，近い将来に実現されるはずの高機能の QA システムなら，この答えを 1 位に欲しいところである．

1. 日本で二番目に高い山は？』と聞かれても答えら
2. 日本で二番目に高い山はどこ
3. 日本で二番目に高い山は？』と
4. 日本で二番目に高い山は？答え
5. 日本で二番目に高い山はどこですか
6. 日本で二番目に高い山は北岳

図 -9 「日本で二番目に高い山は」に対する Kiwi の検索結果

おわりに

これからのディジタルコンテンツの処理を考えるとき，以上に紹介した研究例のように，データ工学・DBMS，マルチメディア情報処理，高速大容量データ処理技術などを踏まえながら，自然言語処理や Web の現実に対応した処理技術や利用技術が求められているということは言をまたない．また，コンテンツの多様性に応じた社会的ニーズに応えると同時に潜在的ニーズを引き出すことで望ましい情報化社会に貢献することが，情報処理技術・システムの最終的なミッションであることも当然である．

こうした問題意識の共有の上で，2001 年度から 5 年間に渡り，A02 には 20 以上の研究グループが参画し，互いの持つ問題意識を確認しながらディジタルコンテンツの持つ今日的課題に学術的観点から取り組んできたわけであるが，決してすべての課題をカバーしていたわけではない．この 5 年間にも，Web の持つ社会的な重要度はきわめて高くなるとともに，さまざまなビジネス展開を狙う企業が新たな試みを提供してきた．当柱の研究は，このような社会のさまざまなセクタの活動と不可避的に関係している．

このような視点で，科研「情報学」の後続の特定領域研究を企画してきた．幸い，2005 年 7 月に新たに平成 17 年度発足特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」（略称「情報爆発 IT 基盤」，領域代表：喜連川優）を発足させることができた．現在進行している，この後続の科研においては，爆発する情報に対応していくための大量情報を管理・融合・活用するコンテンツ処理技術のみならず，安全で堅牢なシステム基盤技術，大量情報を受け止め活用する人間との豊かな共生を実現する技術，知識社会形成とガバナンスに関する技術など，総合的に取り組むことになっている．

この科研「情報学」において，コンテンツに対するさまざまな観点からの大規模共同研究を行えたのは，大変幸運であったと感謝している．この特定領域研究を立ち上げ指揮してきた領域代表の安西祐一郎先生や，評価助言委員会の主査としてきわめて有益な助言をいただいた米澤明憲先生を始め関係者の方々に心から感謝の意を表する次第である．

参考文献

1）http://research.nii.ac.jp/kaken-johogaku/（本科研の報告書や論文等の成果リストが掲載）
2）安西祐一郎（発行責任），安達　淳（編集）：情報学を創る，2006 年 3 月 8 日．（著者に連絡いただければ送付します）
3）大塚真吾，豊田正史，喜連川優：大域ウェブアクセスログを用いた関連語の発見法に関する一考察，情報処理学会論文誌データベース (TOD), Vol.46, No. SIG8 (TOD 26), pp.82-92 (June 2005).
4）Toyoda, M. and Kitsuregawa, M. : A System for Visualizing and Analyzing the Evolution of the Web with a Time Series of Graphs, Proc. of HT2005 - Sixteenth ACM Conference on Hypertext and Hypermedia, pp.151-160 (Sep. 2005).
5）Tanaka-Ishii, K. and Nakagawa, H. : A Multilingual Usage Consultation Tool based on Internet Searching - More than Search Engine, Less than QA, The 14th International World Wide Web Conference (WWW2005) pp.363-371, Chiba, Japan (May 2005).

（平成 19 年 1 月 12 日受付）

● 安達　淳（正会員）
adachi@nii.ac.jp
1981 年東京大学大学院工学系研究科博士課程修了．工学博士．現在国立情報学研究所教授．東京大学大学院情報理工学研究科教授を併任．情報検索，電子図書館システム等の開発研究に従事．

● 喜連川優（正会員）
kitsure@tkl.iis.u-tokyo.ac.jp
1983 年東京大学工学系研究科情報工学専攻博士課程修了．工学博士．同年，同大生産技術研究所講師，現在，同教授，戦略情報融合国際研究センター長．データベース工学，Web マイニングの研究に従事．

● 中川裕志（正会員）
nakagawa@dl.itc.u-tokyo.ac.jp
1975 年東京大学工学部卒業，1980 年同大学院修了（工学博士）．同年より横浜国立大学勤務．1999 年東京大学情報基盤センター教授．現在に至る．自然言語処理，WWW 等の研究に従事．