テキストからの知識抽出による新しいWeb情報アクセスに向けて

インターネット上には膨大なテキストが存在しています．これらから意味情報を抽出して計算機で利用できる形式に変換できれば，今までにないWeb情報アクセスを実現することができます．本稿ではこの分野への導入を述べるとともに，「リッチインデクシング技術」を中心にした取り組みについて解説します．

菊井 玄一郎（きくい げんいちろう） /松尾 義博（まつお よしひろ）
NTTサイバースペース研究所

テキスト情報アクセスの重要性

Web上には膨大なコンテンツやサービスがあります．これらを整理して人々が必要とする情報へのアクセスを支援することはポータルサービスにとってもっとも重要な機能の1つです．そのニーズはきわめて大きく，2006年時点の調査によると，8 000万人を超える日本のインターネットユーザのうち90％以上が1日1回以上検索サービスを利用しているそうです[(1)]．

Web上の情報へのアクセスを支援するうえでテキストの処理は重要な位置を占めています．いうまでもなく，Web上には日本語だけでも億の単位を超えるといわれる大量のテキストがあり，幅広い内容をカバーしています．特に近年はブログや掲示板のようなCGM（Consumer Generated Media：消費者生成メディア）の普及によって，一般の人々の考えや感想など従来のメディアにはあまり現れなかった情報も増えてきました．大量のWebテキストからこれらの情報をうまく取り出すことは，人々のニーズを満たすうえで極めて重要です．また，画像や音楽などの非テキストコンテンツのアクセスにおいても，キャプションやリンクのかたちで付与されているテキストの情報を取り出すことにより，より的確な支援が可能になります．

そこで本特集ではWeb上のテキストに焦点を当て，これらに含まれる情報へのアクセスを支援するために取り組んでいるテキスト（自然言語）処理技術について紹介します．なお，ポータル技術全般に対する研究開発動向については他文献[(2)]をご参照ください．

テキスト情報アクセスにおける課題

Web上のテキスト情報を探すとき，まず使うのがキーワードによる検索エンジンだと思います．これは，入力されたキーワード文字列を含むWebページを検索して，各社独自の順位付けに従って上位から10件ほどのリストを提示するもので，シンプルさと汎用性の高さから，広く利用されています．

ところが，この方法にもいくつかの問題点があります．

第1に検索漏れや検索結果に含まれるゴミ（検索結果に入れてほしくない情報）の問題があります．例えば，あるスポーツ選手の情報を知りたいとしましょう．苗字だけで検索すると同姓の人物が大量に検索されてしまいます．ではフルネームで指定したらどうでしょうか？　同姓同名の人物がWeb上に登場していなければゴミは少なくなるかもしれません．しかし，今度は検索漏れが発生します．テキスト中では同一の人物が「山口選手」のように姓だけで現れたり，ニックネームで現れたりするからです．

第2の問題は検索対象が「情報」ではなく，「文書」であるということです．もちろん文書そのものを探している場合はこれで良いのですが，本当に知りたいことが，人物，店舗，商品などに対するプロフィールや評判などの「モノそのものに関する情報」である場合，これらの事物の名前でキーワード検索し，その結果にいちいちアクセスしてほしい情報を探さなければなりません．また，ある事物に対して多くの人がブログ上でどのように評価しているかを表す「評判情報」などは，その事物に関する書き込みから評判に関する言語表現を抜き出し，一定量以上集めることによって初めて得られるものであり，キーワードともっとも関連しそうな文書を選択することを主眼としている「文書検索」では得られません．

テキストから意味情報の世界へ

以上のような問題を解決するために

は，結局，「テキスト中で個々の言語表現がどういう意味を持っているか」ということを分析し，計算機で扱いやすいかたちで抽出する（例えば，同じ意味を持つ言語表現は同じデータに変換してデータベース化する）ことが必要です．これをあらゆる言語表現に対して行うには基礎レベルからの息の長い研究が必要ですが，私たちは当面のターゲットを実用的な見地から重要性の高い「固有表現」（後述）に絞り込むことで，意味情報を抽出する技術の早期の実用化を目指しています．

■リッチインデクシング技術

「リッチインデクシング技術」とはテキストに出現する個々の固有表現に対して，①それが実世界のどういう事物に対応するか，②テキスト中でどのように言及されているか，といった豊富（リッチ）な情報を付与する技術です．ここで固有表現とは人名，地名，商品名など事物の名前のことで，テキストの意味を考えるうえでキーとなる言語表現です．

リッチインデクシング技術によって付与しようとする情報の例を**図1**に示します．この図の太い四角の枠内が入力テキストで，ピンク色の吹き出しが，自動的に付与される情報です．「アキバ～」の文を例に取ると，「X社」が組織名（会社名）であるということ，「PC-Q」も組織名（会社名）であり，いくつかの支店のうち，このテキストでは外神田にある支店を示していること，書き手は「アフターサービスが良い」と評価していることなどの情報が与えられています．

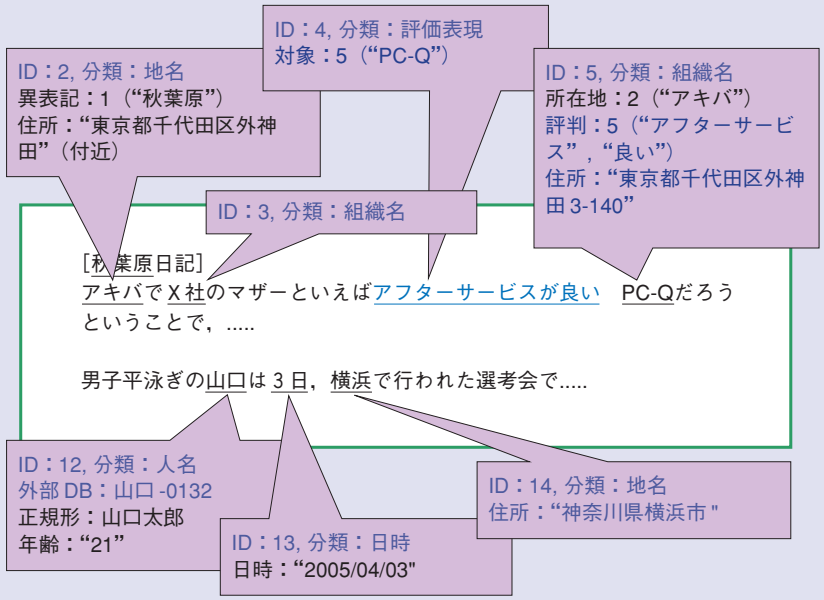

図1　リッチインデクシング技術により付与される情報の例

また2番目の例文では，「山口」が地名ではなく，水泳選手の「山口太郎」という人名の一部であり，人名データベースの「山口-0132」に紐付けられていること，「3日」が実は2005年4月3日であることなどが与えられています．

このような付加情報を与えることにより，テキストで出現しているかたち（文字列）にかかわらず事物そのものの情報を探すことが可能になります．例えば，山口太郎という水泳選手の情報を知りたければ「山口-0132」に紐付けされたテキストを探せば過不足なく見つけることができます．テキスト中の言語表現が実世界の何に対応しているかが分かると，ほかにも応用が広がります．例えば，テキスト中の歌手の名前がCD販売用の商品データベースとリンクされていれば，広告やオンライン販売などと効果的に連携させることができます．

また，付与された情報を表の形式にまとめると，データベースと同じように情報の集計，並べ替え，検索などができるようになります．例えば，ブログごとにどの店舗に対してどういう評判が書き込まれているかを表の形式にまとめると，ある店舗に対してどういう評判が書き込まれているかが分かりますし，各店舗に対する所在地情報と組み合わせると「ある地域で評判のよい店舗のリスト」などを取り出すことができます．

■リッチインデクシングの構成要素

図2に示すようにリッチインデクシング技術は大きく3層の技術に分かれます．

まず，一番下の層は日本語を処理するうえで基盤となる技術で，日本語基本解析技術と語彙知識・オントロジー技術に分けられます．日本語基本解析技術は，入力された日本語文を単語に切り離し，さらに，これらの構文的関係（主語―述語の関係など）を解析します．リッチインデクシングにおいて大きな役割を果たす固有表現もここで取り出します．語彙知識・オントロジー技術は各単語の意味やそれらの間の関係などを扱うための辞書などであり，基礎研究の成果も取り入れられています．

次に，真ん中の層はリッチインデクシング特有の要素技術であり，意味関係抽出技術と固有表現グラウンディング技術から成り立ちます．意味関係抽出技術は各固有表現と意味的に関係のある言語表現を文中から見つけてデータベース化します．固有表現グラウンディング技術は固有表現に対して実世界における事物を対応付ける技術です．

最後に，一番上の層は下の層で得られた情報に基づいてサービスに合わせた知識抽出を行う部分で，ブログなどの口コミテキストから人や物，サービスなどに関する評判を抽出する評判情報インデクシング技術や，CGM中で語られている事物の情報をデータベースのように検索できるCGMマイニング技術などがあります．

なお，これらのうち主なものについては本稿以降の各記事で解説しています．

■リッチインデクシングで広がるサービス

事物の名前が正規化されて，これに関するテキスト中の情報が付与されると，**図3**に示すように，商品やサービスに対するWeb上の評判を分析して提示する「評判検索」や，Webテキスト中の企業活動に関する文章を分析して，例えば「ある商品を共同開発した」といった企業間の関係を抽出する「企業関係マイニング」などさまざまな新しいサービスが可能になります．これらはポータルサービスだけでなく，テキスト情報アクセスに関するシステム開発やASP（Application Service Provider）などのビジネスにも貢献できると考えられます．

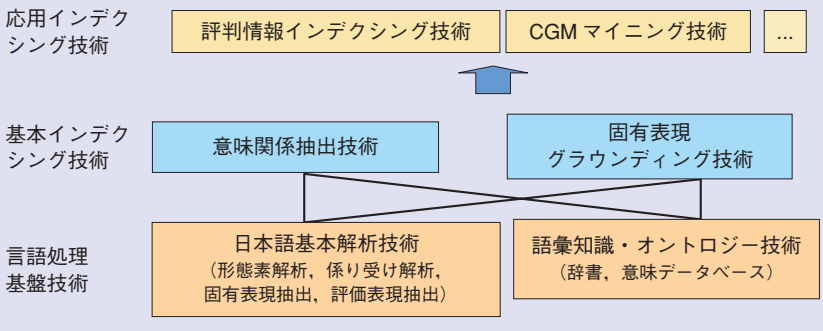

図2　リッチインデクシングを構成する技術

より汎用的な知識抽出技術に向けて

リッチインデクシング技術はWebテキストを知識源として利用する1つの方向性を示していますが，人間がテキストから知識を取り出す能力に比べるとその機能はまだまだ限定的です．より汎用的な意味情報の抽出技術の実現に向けて，NTTコミュニケーション科学基礎研究所を中心にいくつかの試みが行われています．

まず，固有表現に限定せず，一般的な言語表現に対して「何がどうした」といった意味情報を抽出する研究が挙げられます．さらに，これを発展させて，表層に現れていない背後の意味を推定する研究も行われています．例えば，「AがBに勝利した」という文を読むと，私たちは「AとBが対戦した」ということも事実であると理解しますが，このようなことを自動的に行おうというものです．これらの研究については本特集『汎用的な意味解析技術への挑戦』で紹介します．

また，我々人間は新しい言語表現や用法を学習して即座に使いこなせる能力があります．この能力を計算機で実現しようとする研究も精力的に行われ

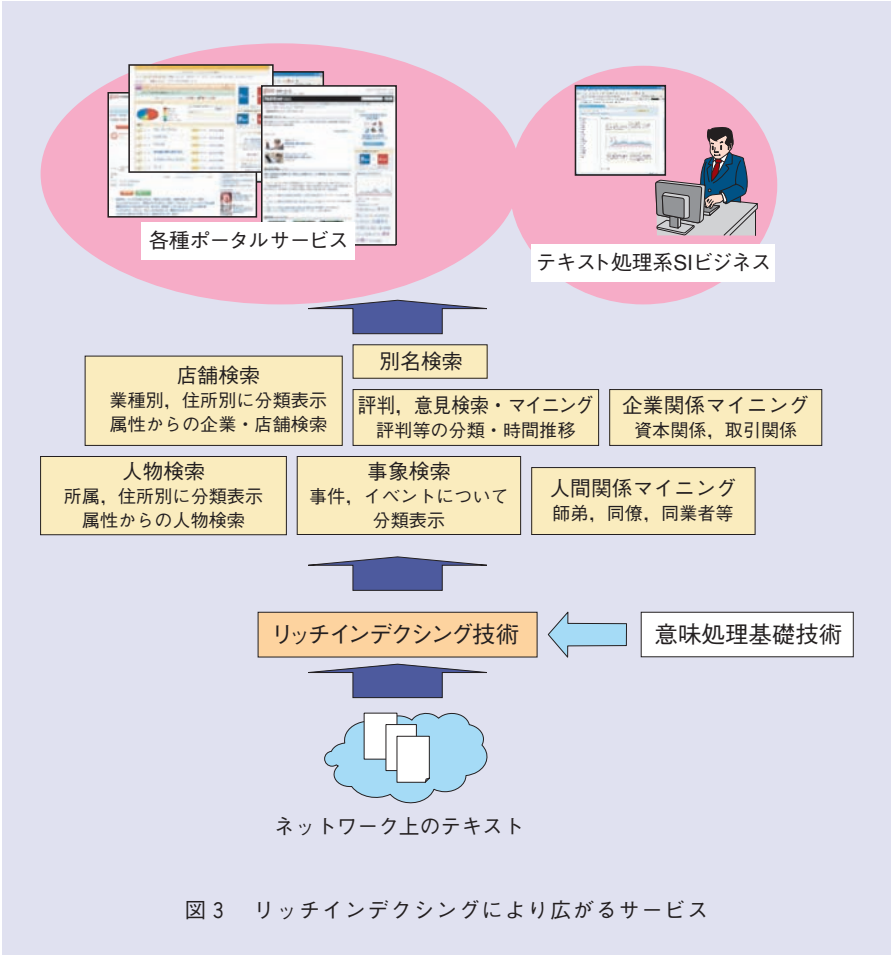

図３　リッチインデクシングにより広がるサービス

ており，その成果の一部はリッチインデクシング技術にも取り入れられています．

おわりに

今まで述べてきた「テキストから意味情報（あるいは知識）を取り出す」という話は，20年ほど前に「人工知能」の世界で扱われた課題であることから，実用には程遠い「夢物語」のように思われる読者もおられるかもしれません．

確かに言葉の背後にある意味を扱うことは，人間の知性の本質に迫らなければならない点で非常に難しい問題であることには変わりありませんが，当時とはいろいろな点で違いがあります．第１に，本稿以降の記事で説明するように，計算機パワーと大規模な言語データベースを駆使した当時と全く異なる方式により，実際のWebテキストやブログなどがかなり精度よく扱えるようになってきたことが挙げられます．第２として，ネットワーク上のテキストはすでに人手で意味情報を抽出する量をはるかに超えており，精度の面で完璧でなくても計算機によって自動的に意味を抽出することにより新しい有益な情報が得られることが挙げられます．

引き続き，テキストの知識化に向けて現実の問題を解決するとともに，よりチャレンジングな目標に向けて研究開発を進めていきたいと思います．

■参考文献

⑴ “インターネット白書2007，” インプレスR&D，2007.
⑵ 特集：“次世代ポータル技術，” NTT技術ジャーナル，Vol.18，No.5，pp.6-31，2006.

（左から）菊井　玄一郎/ 松尾　義博

膨大なテキストから少しでも役に立つ知識を取り出せるよう，研究開発を進めたいと思っています．

◆問い合わせ先
NTTサイバースペース研究所
TEL　046-859-2686
FAX　046-855-1054
E-mail　kikui.genichiro@lab.ntt.co.jp