

ニュースから概念を自動抽出する セマンティックエンジンの研究開発

野村総合研究所
技術調査部 主任研究員

亀津 敦(かめつ あつし)

IT 動向の調査と分析を行う IT アナリスト。専門は、情報系システム全般（主に EIP や検索技術、ナレッジマネジメントなど、情報活用基盤）の技術動向分析と、次世代 Web 技術の知識分野への応用に関する調査研究。



1. セマンティック技術への取り組みの背景.....	17
2. セマンティック技術への社会的なニーズの高まり	20
3. ニュース記事からの概念自動抽出エンジンの研究開発	24

要旨

2008 年から 2009 年にかけて、検索技術の進化に伴って“Google の次にくるもの”として、セマンティック技術を応用したサービスが次々と登場した。その背景には Web2.0 ブームによって質、量ともに拡大したネット上のコンテンツの増加と集合知の拡大がある。また、社会的にも、新聞業界の業界再編や金融業界を襲った経済危機などの影響で、知的なナビゲーションや概念の分析に対するニーズの高まりがある。本稿では、そのような背景の下で NRI（野村総合研究所）技術調査部が研究開発を行った、ニュース記事からの概念自動抽出を行うセマンティックエンジンの研究開発の概要について解説する。

キーワード：セマンティック技術、W3C、検索エンジン、アルゴリズムトレード、RDF、RDF トリプル、オントロジー、SPARQL、OWL、タギング、概念自動抽出、情報抽出エンジン

From year 2008 to year 2009 along the evolution of searching technology, the services applying semantic technology have appeared one after another as "What comes next to Google". The increase of internet contents which under affection of Web2.0 have amplified their quality and quantity and the expansion of collective intelligence are behind it. Moreover, following the influence of restructuring of Newspaper industry and the economic crisis that hit the financial industry etc., rising of the needs for intellectual navigation and conceptual analysis are also there. This report, explains about the outline of the research and development that under the above situation, technology research department of Nomura Research Institute, Ltd. (NRI) has performed on a semantic engine that automatically extracts the concept from the news articles.

Keywords : Semantic technology, W3C, Search engine, Algorithm trade, RDF, RDF triple, Ontology, SPARQL, OWL, Tagging, Concept automatic extraction, Information extraction engine

1. セマンティック技術への取り組みの背景

(1) セマンティックサービスの台頭

① 検索エンジンの進化とセマンティック技術

昨年、2008年から2009年前半にかけてのインターネットサービスの世界でのホットな話題は、「セマンティック技術」の名を冠したサービスが次々と登場したことに集中した。

2008年7月、マイクロソフトがセマンティック検索サービスを手がけるベンチャー企業「Powerset」を買収する、というニュースが報じられた。米国で立ち上がったばかりのベンチャー企業ということもあり、特に日本の一般的なネットユーザーやビジネスパーソンにはなじみのない話題であったが、「マイクロソフトがGoogleに対抗するための買収」と

いう観測や、その買収額の大きさ（公表はされていないが、複数のIT系ブログメディアが報じたところによると1億ドルとも言われる）から、一般の新聞などでも報じられることとなった。

この買収劇に先立つこと数ヶ月、2008年の4月頃から米国ではPowersetのほかにもいくつかの“セマンティック技術”を売りにしたサービスが公開され始めていた。冒頭のPowersetのほか、セマンティック技術を応用し、自動タグ付けを行う「Twine」などのベンチャーが、“Googleの次に来るもの”として次々とベータ版サービスを公開し、Webサービスの表舞台に登場し始めていたのである。

マイクロソフトに買収されたPowersetが



図1 セマンティック検索エンジン「Powerset」のスクリーンショット

出所：PowersetのWebサイト（<http://www.powerset.com/>）

実現していたサービスは、Web ページに含まれる意味を取り出して操作する、セマンティック Web の技術を応用した概念検索であった。セマンティック Web そのものは、既によく知られているように W3C (World Wide Web Consortium) のティモシー・ジョン・バーナーズリーによって 1999 年に提唱され、標準化が進められてきた技術である。用いられている技術自体は決して目新しいものではなかったのだが、Powerset はこのセマンティック Web 技術を Web から“本当に知りたい概念を引き出す”ために応用した。具体的には、ネット上のさまざまな人々によって編集され、成長する Wikipedia の情報をデータソースとして、自然文で入力した質問に対する回答を得たり、知りたいモノや場所、事件などに関して整理、体系化された情報を得たりすることができる。例えば、「Google が買収した企業はどこか?」「FDA (Food and Drug Administration of the United States Department of Health and Human Service : 米国食品医薬品局) が禁止した薬品は何か?」といった質問に対して、該当する具体的なモノを挙げて回答する。

Powerset のような、セマンティック Web の技術を活用した情報検索「セマンティック検索」は、従来のキーワードに該当した結果が一覧表示される検索エンジンと異なり、利用者が本当に必要とする情報を引き出すという点で大きな進化を果たしている。このため、Powerset は

“Google キラー”と呼ばれ、マイクロソフトによる買収につながったのである。

Powerset のサービスは買収後、2009 年 5 月に公開されたマイクロソフトの新しい検索サービス「Bing」に統合され、Google、マイクロソフト、さらに Yahoo をも巻き込む検索サービスのトッププレイヤー達の競争に影響を与えた。マイクロソフトの Bing 公開の 2 ヶ月前、Google は同社の検索エンジンにセマンティック技術を採用したと発表した。具体的には、検索結果の一覧に加えて、ユーザーが検索したキーワードと意味的に関係の深い単語を、キーワード絞り込みの候補として表示するようになったのである。

2008 年から始まった、検索サービスへのセマンティック技術の応用は着実に広がり始め、インターネットからの情報の発見と活用を進化させ始めている。

② セマンティック技術への注目の要因

これらの“Google の次”を標榜するサービスがこの時期に登場した背景は、2006 年初頭から始まった Web2.0 ブームと深い関係がある。

Read only な Web と対比される、Web2.0 以前と比べて、Web2.0 時代の Web コンテンツは質・量ともに爆発的に拡大した。Powerset も利用している Wikipedia が典型例であるが、インターネットの利用者が一方的にコンテンツを読むだけでなく情報を加工し、ブログや SNS (Social Networking Service) を利用して新たな

情報を発信するということが行われるようになった。また、利用者がネット上の情報に分類キーワードをつけて共有する“タギング”という文化が定着した。

これらがもたらしたものは、Webのページに書かれたコンテンツが、単なるテキストのつらなりから、半構造化された「集合知」へと進化していくという現象である。前出のPowersetが質問文に的確に回答することができるのは、Wikipediaという、ネットの多くの人々が知恵と手間を持ち寄って情報を整理したデータをデータソースとして利用しているからである。また、タギングによって、ある言葉がどのような意味上の分類に当たるのか、他の言葉とどのような関係にあるのかを人々が自発的に教えてくれる結果を利用することができるようになり、単なるキーワード検索以上の概念の整理や発見がしやすくなったことに負うところが大きい。

1999年からW3Cで標準化が進められてきたセマンティックWeb技術が、昨年から注目を集めるまでなかなか表舞台で脚光を浴びるに至らなかった理由もここにある。W3Cで標準化が進められてきたセマンティックWeb関連の仕様は、1) ある概念を、単なる単語としてのテキストとして表現するのではなく、関係する情報(分類や、他の概念との関係など)を含めてどのようにマークアップし、操作するか(RDF、SPARQL関連)、2) 概念と概念のつながりをどのように表現し、コン

ピューターがコンテキストを理解できるようにするか(OWLなどオントロジー表現関連)、3) 2) で表現された概念と概念のつながりを元に、いかにコンピューターに人間のように推論をさせるか、という段階で進んできた(図2)。

これらの標準は、仕様としては2004年にはベースとなるRDF、オントロジー表現関連については仕様策定がなされており、つい最近できたものというわけではない。それが昨今改めて注目を集めているのは、Web2.0ブームに始まったインターネット上でのコンテンツの拡大と“集合知の登場”によって、仕様策定された概念の表現を実際に利用できるコンテンツと知識の抽出が現実味を帯びてきたからである。

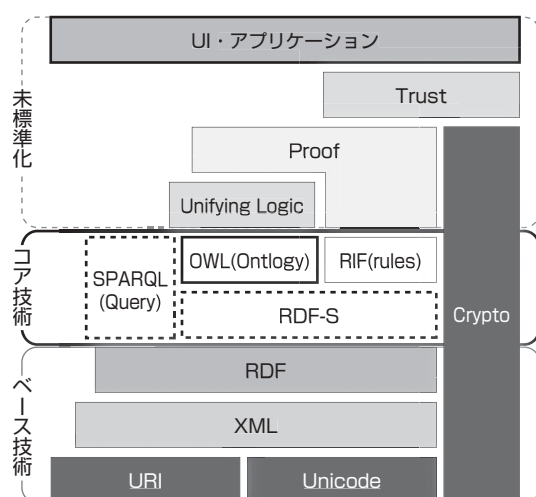


図2 セマンティックWeb関連の仕様のレイヤー構造
出所: W3Cの標準化案を基にNRI作成

2. セマンティック技術への社会的な ニーズの高まり

これまで振り返ってきたセマンティック Web 技術への注目は、Web2.0 ブームに始まる技術と環境進化に後押しされたものであった。そういう意味では、技術主導での新サービスの開始という性格が強い。それでは、実際にセマンティック Web 技術を利用するユーザー企業や産業界のニーズは盛り上がっているかが気になるところである。

米国においては、産業界の側でもやはり昨年から今年に入って、セマンティック Web 技術への期待が高まっている傾向が見え始めた。

その典型例として、米国での新聞業界と金融業界の動向を見てみたい。

(1) 知的ナビゲーションの装備に向かう

米国新聞業界

インターネットの進化に伴うコンテンツのデジタル化、メディア利用傾向の変化の影響

を受けて新聞業界が厳しい状況に置かれているのは先進国ではどこでも共通の状況であるが、全米に 1,000 紙以上の日刊紙があるといわれている米国の新聞業界では特に売上の落ち込みが早く、老舗の新聞社による新聞の廃刊や紙媒体による発行の廃止、オンラインへの移行が報じられ始めている。

紙媒体での苦境と対比的に、米国の新聞社大手はオンライン版での記事の検索機能の強化に努めている。

図4は、ワシントンポスト紙の記事の画面である。ワシントンポストのオンライン版では、一般の記事を表示しているページの下部に、図4右側のような、概念と他の概念をつないだネットワーク図のようなビューが表示される。これは、同紙が採用した Evri 社の「セマンティックコンテンツ発見エンジン」が記事を意味的に解析し、関連がある概念や他のページを表示するものである。

図の記事の例 (Apple 社の新しい iPod nano

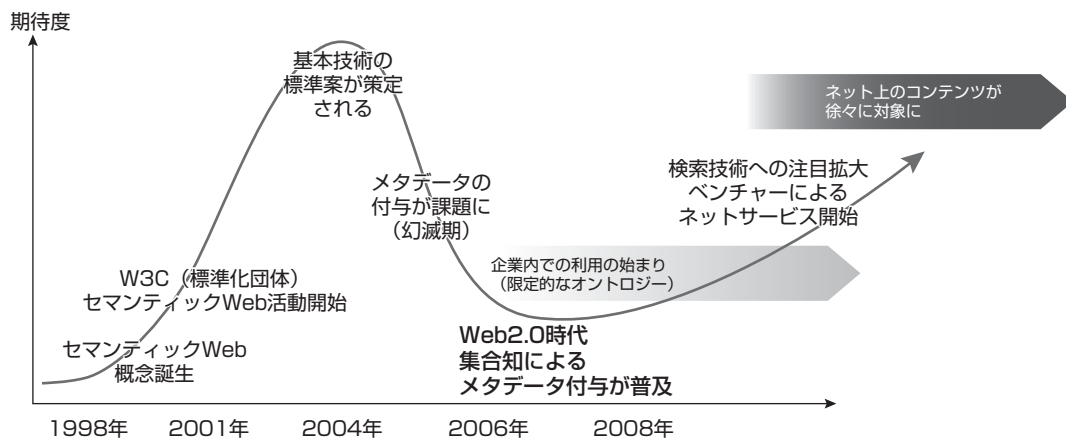


図3 セマンティック Web への期待の変遷

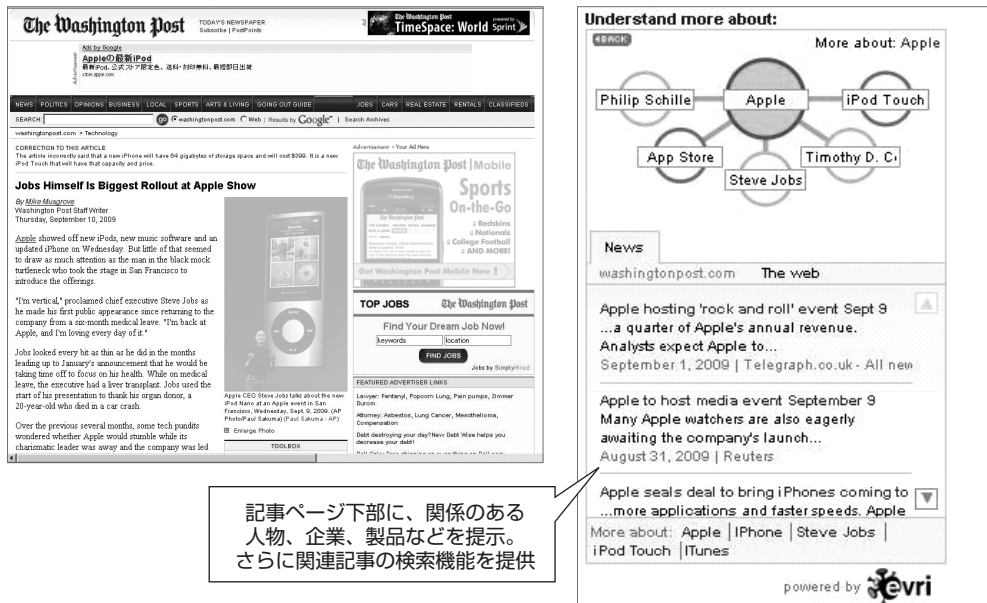


図4 ワシントンポストが採用したセマンティックコンテンツ分析エンジン「Evri」のスクリーンショット
出所：ワシントンポストのWebサイト
(<http://www.washingtonpost.com/> 要ユーザー登録、ログイン)

の発売に関する記事)では、記事中に含まれる会社名や人名、サービス名、商品名を概念として識別し、Appleという企業名とリンクさせている。さらに、それぞれの概念を同じように含む他の記事をリストアップする。

オンライン新聞サイトでこのような仕組みを用意することで、ある記事を読んだ読者を、ほかにも興味を持ちそうな記事に誘導することができる可能性が高まる。従来のキーワードの類似頻度だけで“関連記事”を推奨するのとは異なり、“リンゴを意味する Apple”を含む記事をコンテキストを誤って表示させることはない。また、「Apple社＝音楽プレーヤーのメーカーである」という概念のつながりを

解析することで、キーワードには現れなくても、競合企業の名前が入った記事や、DAP (Digital Audio Player：デジタルオーディオプレーヤー) 市場に関する記事を推奨する、といったことも可能だ。

新聞社としては、的確な関連記事へのナビゲーションを行うことで、ページビューの拡大やサイト滞在時間の増大を期待することができ、カスタマーロイヤルティの維持とオンライン広告収入の拡大につなげることができるだろう。また、図5はフィナンシャルタイムズが提供しているセマンティック分析／検索画面である。この例でも同様に、Apple Computer, Inc.という与えられた企業名を元

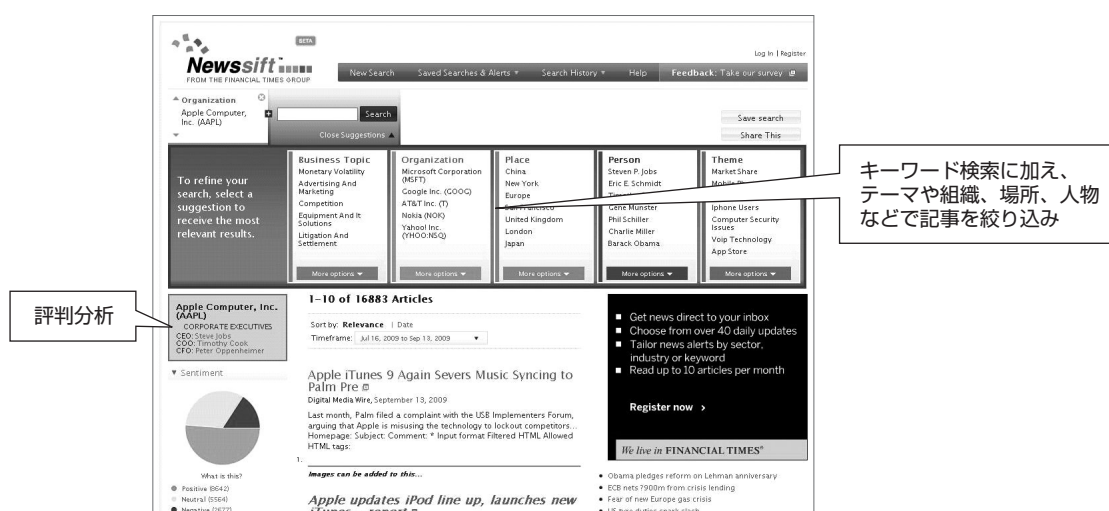


図5 フィナンシャルタイムズのセマンティック検索「Newssift」のスクリーンショット
出所：NewssiftのWeb サイト (<http://www.newssift.com/>)

に、関連する記事をビジネスピック（広告、マーケティングや市場での競合など）に関する記事や他の組織との関係を含む記事、場所、人といった概念ごとに分類して絞り込むことを可能にしている。今後、オンラインコンテンツの蓄積が進み、記事データベースとしての価値が増すにつれてこのようなセマンティック検索機能の存在は非常に重要になってくるだろう。

（２）概念の分析を活用し始めた金融業界及び金融情報サービス

リーマンショックに端を発する金融危機以降、金融業界では市場分析を行う上で大きな変化が生じている。証券アナリストやトレーダー、ファンドマネージャーなどの投資の意志決定にかかわる人々が市場を読むに当たっ

て、金融危機以前のようなテクニカル分析による予測が当たりにくくなっているというのである。これは、金融危機勃発以降、市場が非常事態のような状態に置かれているため、規制当局の打ち出す施策や政府の動向など、外部要因によってマーケットが大きく変動することが原因である。このような状況下では、時々刻々と入ってくるニュースが意志決定にこれまで以上に影響力を持つようになる。

また、金融危機前後から始まった投資商品の多様化も、ニュース記事などからの情報の発見と意志決定への活用の重要性を増す要因となっている。金融危機顕在化前や、最近の原油価格高騰に見られるように、投資資金が向かう先に、証券市場だけでなくこれまでなじみのなかった商品市場も入ってくるようになっており、判断のためのニュースや解説記事を多く読むことが

必要になっているという。

こうしたことから、主な金融情報ベンダーの多くはセマンティック技術への取り組みを以前にも増して強化しており、ニュース記事をいかに速く活用できるサービスとして提供できるかを模索している。

ロイターは2007年4月にセマンティック技術によるテキスト分析ソリューションを提供する Clear Forest 社を買収して主に金融業界向けのセマンティック分析ソリューションを提供している。

同社のソリューションでは、ロイター、ブルームバーグなどの主要な金融ニュースのテキストを入力データとし、種々の自然言語解析技術や辞書を活用して記事に含まれる概念（企業名や日付、人名、出来事など）を取り出す。取り出された概念を元に記事にメタデータを付与することで、ニュース記事に対する知的な処理を行うアプリケーションの構築を可能にする。

例えば、ニュース記事の中に特定の企業名が含まれている場合、その企業の株をポートフォリオに組み込んでいるファンドマネージャーにのみアラートを発して最新情報が入ったことを伝える、といったリアルタイム性の高い意志決定支援システムが想定されている。また、過去のニュース記事を大量にバッチ処理で読み込んで概念の自動抽出を行い、そのニュースとニュースが発生した時点の過去の株価データとをつきあわせることで、金

融ニュースが株価のパフォーマンスにどのような影響を及ぼすかを分析する、といった利用も想定されている。

これは、市場に流れるニュースをアルゴリズムトレードシステムによる自動執行のトリガーの1つとして活用する、という目的で利用される。

米国の金融市場ではこのようなニュースに含まれるテキスト情報をアルゴリズムトレードに活用している例が増え始めていると見られている。その傾向が2008年9月の米国の金融マーケットのある事件で垣間見られた。

2008年9月8日に、ユナイテッド航空の親会社である UAL (United AirLines) の株価が12.50ドルから3ドルへと急落した。唐突に始まったこの値動きのきっかけは、「UALが破産を申請」という記事がGoogle ニュースに載り、ブルームバーグによって機関投資家などに配信されたことによる。

実は、元になった UAL の破産申請というニュースは2002年のニュースで、過去のニュースが誤って広まったものであった。なんらかの理由によって米国のある新聞社のサイトの“アクセス上位の記事”に2002年の問題の記事がランクインし、トップページにリンクが浮上したものをGoogleのクローラが発見してGoogle ニュースに掲載した結果広がってしまったのである。

元の過去記事には日付がなかったが、人間が読めば明らかに“誤報”とわかる記事であ

ったという。にもかかわらず、ニュースが流れたとたんに株価が急落したことで、ニュースのキーワードを用いた自動取引が多くなされていることが伺える結果とみられている。

この事例は、ニュースをシステムが誤って解釈したがゆえに望ましくない結果となった例ではあるが、少なくとも2つの示唆を見取ることができる。

- 米国の金融業会では、既にテキストベースのニュースを迅速な意思決定に活用し始めているということ
- ニュースを意思決定に活用するためには、元の記事から適切な概念を正しく抽出することが肝要であること

2点目は今後のセマンティック技術を活用したアプリケーションの構築に際して、特に重要である。Clear Forestのソリューションでも、競争力のキモとなるのは、自然言語で書かれたテキストの中から、いかに適切に概念を抽出できるか、という能力である。

セマンティック技術の研究開発において、概念のマークアップや関係性の表現（知識分野の概念の体系、すなわちオントロジー）の仕様策定、構築についてはかねてより検討が重ねられてきた。

それに加えて、これからのホットトピックとして、昨年頃からまさにこのテキストからの概念の自動抽出「Entity Extraction」をい

かに実現するかに注目が集まり始めている。

3. ニュース記事からの概念自動抽出エンジンの研究開発

（1）概念自動抽出エンジンの研究開発の目的

これまで見てきたような時代背景と技術への注目トレンドを検討した結果、NRI技術調査部でも、テキストからの自動抽出エンジンの研究開発を行った。

我々は、今後ネット上に大量のデジタルデータがあふれる時代を見据え、セマンティック技術を活用して人間の情報の整理、発見や意志決定をいかに速くするか、また、コンピューターに大量の概念の関係を分析させることで、人間だけでは見つけられなかった情報群からの新たな気づきを得ることを目的とした。その端緒として、ビジネスニュースに含まれる企業の活動の概念自動抽出を行い、企業のサービスや研究開発、資本移動などに関する検索や分析を行うアプリケーションの実装を行った。

（2）構築したシステムの概要

本研究開発のメインとなる概念自動抽出エンジンを含む、システム全体の概要は図6のようになっている。

まず、実験データとして用意した新聞記事のテキストデータが、記事からの概念自動抽出を行う概念自動抽出エンジン「Smart Collector」に投入される。Smart Collectorは、

個々の新聞記事のテキストデータから、以下のようなデータを抽出する。

- 企業（組織）名
 - 日付
 - 企業活動
 - └ 研究開発
 - └ 商品の開発
 - └ 商品の販売
 - └ 提携（資本提携、業務提携…）
 - 商品／サービス名
- など

ここで抽出されるべき対象の概念は、企業の活動を記述するために必要な概念を研究者がリストアップし、それぞれの概念同士の体系を定義した。我々はこれを「企業オントロ

ジー」と呼んでいる。

個々の記事は、上記の企業オントロジーを構成する概念が含まれるかどうか、情報抽出エンジンの Smart Collector が解析を行う。例えば、「A 社は、B 社と提携し、ハイブリッド車向けの燃料電池の製造を開始した。」という記事からは、次のような情報が抽出され、企業オントロジーに照らし合わされ、関連する情報同士が組み合わされてデータベースに蓄積される。

このような、概念と概念とその関係の3つの要素を組み合わせたセットは、セマンティック Web 関連の仕様で策定されている RDF を用いて表されるため、「RDF トリプル」と呼ばれる。

個々のニュースが読み込まれ、解析されるたびに抽出された RDF トリプルが抽出され、

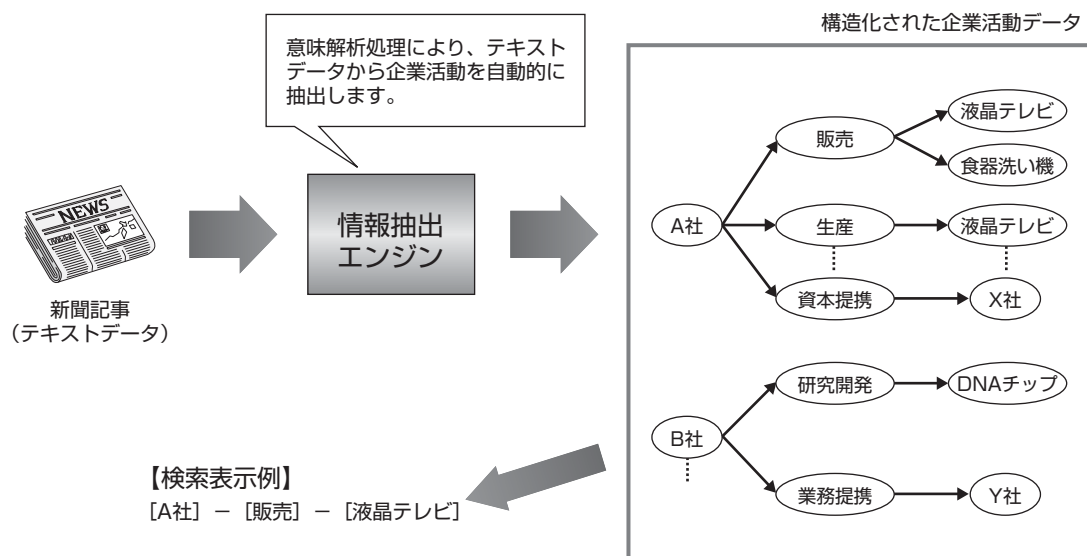


図6 NRI が今回研究開発した概念自動抽出システムの全体像

とができる。図7は、燃料電池を研究開発している企業」のみを抽出した例である。通常のキーワード検索と違い、単にキーワードが入っているだけで記事がヒットすることではなく、本来探したい「(企業名) —研究開発→燃料電池」というファクトを含む記事だけを表示することができるようになり、情報検索の効率化に寄与することができる。

反対に、ある企業に着目して、その企業が行った他社との提携や企業買収、研究開発しているテーマや販売を開始した製品名などを一望するようなことも可能である（図8）。このような視点からのニュースへのアプローチは、マーケティング担当者やアナリストなどのような、企業や業界動向を分析する立場の人々の知的活動の効率化に有効であると考えられる。

このように蓄積されたデータベースを活用することで、ユーザーは企業活動に関する概念を指定して、必要な情報のみを検索するこ

企業活動を条件として指定

該当する企業名と企業活動を表示

図7 蓄積された企業活動の検索例（「燃料電池を研究開発している企業」のみを抽出）

(3) 今後の自動分類エンジンの発展・応用の方向性

今回 NRI が開発した情報抽出エンジンは、自然言語処理技術とセマンティック技術を活用することによって、ニュース記事の中から意味的に必要な情報を迅速に取り出すことを実現可能なことを示した。日本語のニュース記事についても、欧米の新聞社が行っているような意味に基づいた関連情報の提示や、過去の記事データベースを分析的に活用することにセマンティック技術を応用し、より高い付加価値をユーザーに提供したり、新サービスを開発したりすることが可能であろう。

今回研究開発した情報抽出エンジン自体は、どのようなテーマを対象とした文章であっても汎用的に利用できることを意識して作成している。そのため、まずは企業活動から

の情報抽出に絞り、データソースをニュース記事とし、オントロジーとして企業の活動を抽象化した企業オントロジーの構築に成功した。同様に、データソースとオントロジーを別のテーマにあわせて変更することで、まったく違った分野の情報抽出も可能になる。例えば、医療関連の症例や薬品の効能などを記述した文書をデータソースとし、ライフサイエンス分野のオントロジーを用意することで、ヘルスケア分野で利用可能な情報抽出と検索アプリケーションを実現することも考えられる。

今後、本技術の応用先を模索し、企業内における情報整理や分析業務の高度化を支援する研究に発展させる予定である。



図8 新聞記事データから個別企業の活動一覧を活動の種類ごとに整理