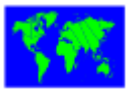


◇KDDI総研R&A　2012年8月号

大変革期に突入する検索エンジン：
「AI（人工知能）」の導入と「ビッグ・データ」への対応

執筆者　KDDI総研 特別研究員　小林雅一

記事のポイント

サマリー

私たちが様々な情報を入手するための基本的な手段である検索エンジンが今、歴史的とも言える大変革期に突入しつつある。たとえばGoogleが2012年5月にリリースした「Knowledge Graph」など新型検索エンジンでは、従来のような検索キーワードに関連したホーム・ページへのリンクのみならず、ユーザーが求めている情報をズバリと画面に表示する。さらに、その先には単なるキーワードではなく、ユーザーの発する質問を理解して、その答を表示する「セマンティック検索」が控えている。

これは自然言語処理と呼ばれ、いわゆる「AI（人工知能）」の代表的なケースである。このAIは最近、一大ブームを形成しつつある「ビッグ・データ」を解析するために、最も有効な手段と目されている。Googleが開発中の「セマンティック検索」や、Appleが先ごろ商品化した音声アシスタント「Siri」などは、このAI技術を駆使してビッグ・データから経済的な価値を引き出す動きと見ることができる。

主な登場者　Google　Wolfram Research　Metaweb Technologies　Facebook　Apple

キーワード　Search Engine　検索エンジン　Knowledge Graph　Semantic Search　セマンティック検索　WolframAlpha　Freebase　構造化データ　structured data　非構造化データ　unstructured data　Siri

地　域　世界　米国

Title	Search Engines Entering an Age of Revolution: Introducing AI and Adapting for Big Data
Author	Masakazu Kobayashi (Research Fellow, KDDI Research Institute)
Abstract	The search-engine, our primary tool to access the wealth of information on the Internet, is now entering a potentially epoch-making age of revolution. For example, the “Knowledge Engine,” which is the foundation of Google’s new search-engine, will be able to give us the exact information we desire, in addition to the normal search-engine function of delivering the webpage links related to the entered keywords. Further, in the near future, Google is expected to release the “Semantic Search,” which is designed to understand our query literally, and provide the user with an exact answer. Collectively, these developments are being referred to as “Natural Language Processing” (NLP), and are considered one of the major and essential parts of Artificial Intelligence (AI). AI is being looked to as the most effective way to analyze so-called “Big Data,” a recent buzzword in industry circles. Google’s aforementioned Semantic Search, or Apple’s high profile Voice Assistant, “Siri,” are at the forefront of a broad-based movement trying to draw some economic value from Big Data.
Keyword	search-engine Knowledge Graph Semantic Search WolframAlpha Freebase structured data unstructured data Siri

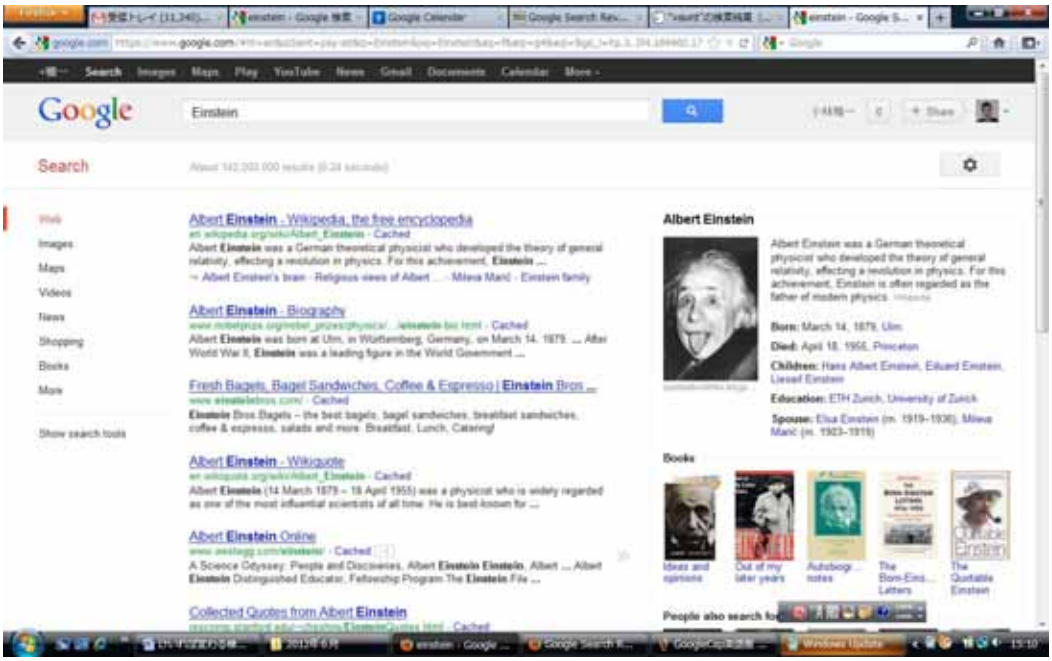
1　Googleの新型検索サービスは、知りたい情報をピンポイントで表示する

Googleは2012年5月中旬、同社の検索エンジン・サービスを大幅に改良した。そこで新たに導入された「Knowledge Graph」などの技術により、検索の対象となる「人」、「モノ」、「場所」などの詳しい情報をピンポイントで表示できるようになった。

この新型検索は今のところGoogleの英語サイト(www.google.com)でしか使えない。日本語サイト(www.google.co.jp)を始め、他の言語サイトで新型検索が利用可能になる時期を、同社はまだ明らかにしていない。しかし逆に、この時差のお陰で、現時点で、例えば日本語と英語サイトの検索結果を比べることによって、Googleの新型検索がどんな姿であるかが、一目瞭然で分かる。それを試みたのが、以下の【図1】、【図2】である。

【図1】従来型Google（日本語サイト）では検索対象の関連HPのリストが表示されるだけ

【図2】新型Google（英語サイト）では検索対象の詳しい情報が、関連HPリストの右側に表示される



ここでは「Einstein」という検索キーワードを、日本語サイトと英語サイトのそれぞれで入力してみた。日本語サイト（つまり新たな技術を導入する前のGoogle検索）では、従来のようにAlbert Einsteinの関連ホーム・ページを列挙（リスト）しているだけだ【図1】。これに対し、新技術を導入した英語サイトでは、関連ホーム・ページのリストの右隣に、Einsteinの個人情報が詳しく紹介されている【図2】。そこにはEinsteinの顔写真、略歴、生・没年月日、家族構成、学歴など、かなり詳しい情報が含まれている（これをGoogleは「Knowledge Graph」と呼んでいる）。

本来、私たちユーザーが何かを検索する場合、従来のような関連ホーム・ページのリストよりも、むしろ検索対象の情報を一発で知りたいと望んでいるはずだ。この点で、「Knowledge Graph」を導入したGoogleの新型検索は、そうした検索の理想に一歩近づいた格好になる。しかし、当のGoogle側では「今回のアップデートは（この先に控えている大変革に向けた）小さな一歩に過ぎない」としている。では、それは今後、どのような姿へと変貌を遂げるのだろうか？

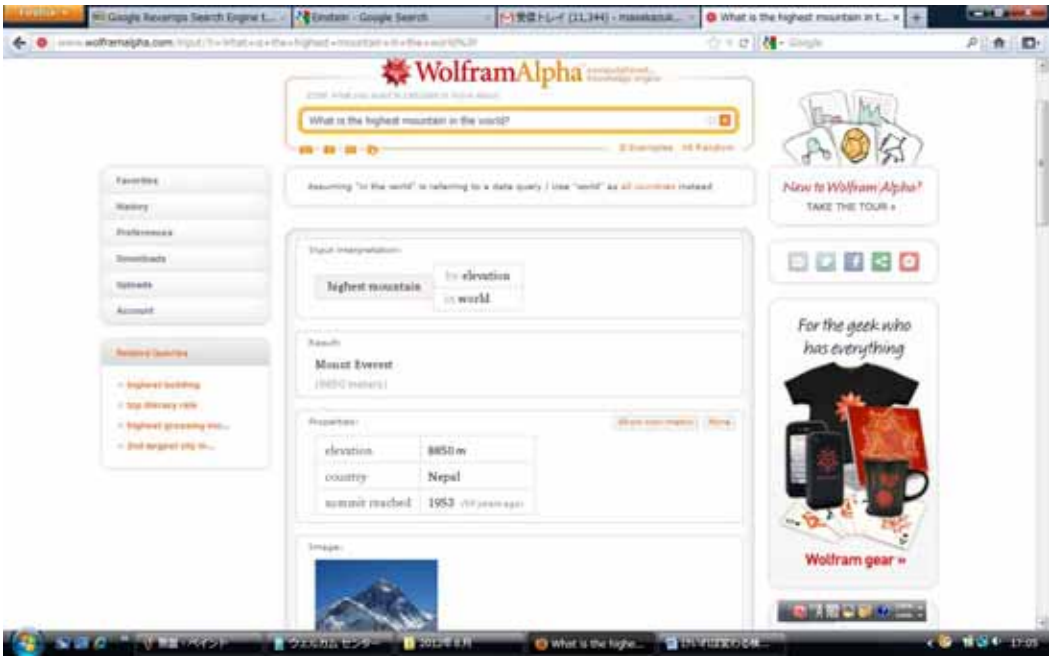
2　次世代検索エンジンは「自然言語処理」で人間の質問を理解する

Google検索の将来像を占う上で参考になるのが、WolframAlpha（http://www.wolframalpha.com/）という新種の検索エンジンだ。前述の通り、従来

型Googleに代表される検索エンジンの場合、ユーザーが検索窓にキーワードを入力すると、それに関するウエブ・サイト（ホーム・ページ）へのリンクがパソコン画面にずらずらと表示・羅列される。

これに対し、WolframAlphaでは単なるキーワードではなく、たとえば「What is the highest mountain in the world?（世界最高峰の山は何ですか？）」など、自分が知りたいことを文章（質問）として入力する。すると検索エンジンの方でも単なるホーム・ページの羅列ではなく、「Mount Everest（エベレスト山）」のように、答え自体をズバリと返して来る。しかも、そこではエベレストの標高、それがある国、初登頂の年、写真、地図上の場所など、関連情報が整理されたパッケージ・データとして表示される【図3】。

【図3】WolframAlphaでは自然言語（文章による質問形式）で情報検索できる



他にも筆者が試したところでは、たとえば「Who is the fastest runner in the world?（世界最速のランナーは誰か？）」「Who is the prime minister of Japan?（日本の首相は誰か？）」など初歩的な質問には、かなり精度の高い情報を返してくれた。

Wolfram Alphaは英国人の数理物理学者、Stephan Wolfram氏が中心になって開発した検索エンジンで、同氏が1980年代に開発したビジュアル指向の数式処理ソフト「Mathematica」をベースにしている。WolframAlphaの最大の特徴は、前述のように「人の言葉（質問）を理解して、それに対する答えを一発で返す」という点にあ

る。それはまた、一種の推論機能も備えているので、質問が多少曖昧で分かり難かったり、文法的に間違っていても、ユーザーの意図をくみとって正しい回答を返してくれる（もちろん回答が間違っているときもある）。

しかし、このWolframAlpha、残念ながら、まだ日本語に対応していない。それでも敢えて、ここに紹介したのは、Googleが現在開発中の次世代検索エンジンが、これと非常に良く似たものになると見られるからだ。2012年3月15日付けの米Wall Street Journal紙に掲載された「Google Gives Search a Refresh」という記事によれば、Googleの次世代検索エンジンは「言葉の意味を理解するプロセス」に基づいており、「Semantic Search（セマンティック検索、意味検索）」と呼ばれる。

もちろん、Googleは従来のキーワード入力型の検索エンジンから、新たなセマンティック検索へと完全に切り替えてしまうわけではない。むしろ従来型検索エンジンのプラス・アルファとして、セマンティック検索を追加する。たとえば検索窓から「世界最速のランナーは誰ですか」と入力したとすれば、結果画面のトップには「ウサイン・ボルト」という名前と、彼の写真や保有する世界記録など関連情報が詳しく表示される（この辺りまではWolfram Alphaと同じだ）。しかし、その近くには従来型検索のように、ボルトについて書かれたホーム・ページへのリンクがずらずらと表示される。ちょうど、そんな形になるのではないか、と上記Wall Street Journalの記事では予想している。

実際、今年の5月にリリースされたGoogleの新型検索では、前述のKnowledge Graphを構成する詳しい情報が、まさにそれに該当する。ただし現時点では、まだWolframAlphaのような自然言語処理は導入されていない。Wall Street Journalの予想に従えば、いずれこの自然言語処理の部分もGoogleの次世代検索に実装されるはずだ。

また、MicrosoftのBingもAI検索のPowersetの技術を取り込むなどの工夫により、Googleと同じ方向を見据えている。

3　構造化データと非構造化データの両方をカバー

これら新種の検索エンジンの内部構造にまで、ちょっと踏み込んで解説すると、実はWolframAlphaもGoogleのセマンティック検索も、従来のウエブ検索と根本的に異なる。特にWolframAlphaの場合、そもそもの検索対象がウエブではない。むしろ、この検索エンジンを開発・提供しているWolfram Researchという会社が管理するデータベースから、各種情報を引っ張ってきて、それをユーザーに提供しているのである。従って、厳密にはWolframAlphaを「検索エンジン」に分類するのは正しくないかもしれない。なぜなら通常、我々が「検索エンジン」と言うときには、暗黙のうちに「ウエブから情報を検索するためのツール」であることを前提としているからだ。

このWolfram Researchが管理するデータベースは、同社が企業や政府機関から購入したり、ライセンス（使用権）を借り受けたものだ。つまり相応のコストがかかっているので、WolframAlphaは基本的に有料サービスである（試しに最初の数回は無料で使えるが、やがて有料サービスへの加入を促すメッセージがパソコン画面に表示される）。

この種のデータベースは、専門的な用語では「構造化データ（structured data）」と呼ばれる。つまり生の情報を加工し、各種の属性に基づいて整理・体系化したデータである（企業などで使われる「リレーショナル・データベース」は構造化データの代表とされる）。これに対し、ウエブ上に存在する無数のホーム・ページは、生の情報がほぼ無造作に置かれた状態であることから「非構造化データ（unstructured data）」と呼ばれる（ホーム・ページは若干の構造性を備えたHTML文法に従うことから「半構造化データ」と呼ばれることもあるが、一般的には「非構造化データ」に分類されることが多い）。要するに従来の検索エンジンがウエブという非構造化データを相手にしていたのに対し、WolframAlphaは構造化データを相手にしている。だからこそ、人間の発する言葉に対応して、正確な答えをズバリと返すことができるのだ。

Googleが新たに提供するセマンティック検索も、基本的にはWolframAlphaと同じく構造化データ（つまりGoogleが構築した自社のデータベース）から情報を引き出してくる。Googleはこれに備えて、2010年に米Metaweb Technologies社を買収し、彼らが当時所有していた1200万件にも及ぶ情報（構造化データ）を蓄積したデータベース「Freebase」を取得。この上にGoogle自身が情報を追加し、現時点で2億件以上の構造化データからなるデータベースが構築されているという（上記Wall Street Journal記事より）。実際、現時点のGoogle検索（英語版のみ）でも、前述のKnowledge Graphの部分は、この構造化データから答えを見つけてきた結果だ。

要するに（まだ予想の域を出ないが）、近い将来始まるGoogleの次世代検索では、セマンティック検索の部分は同社自身が保有するデータベース（構造化データ）から情報を引き出してくる。その下に表示される関連ホーム・ページへのリンクは、従来のようにウエブ（非構造化データ）から検索してくる。こんな形になりそうだ。

ただし圧倒的な技術力を誇るGoogleだけに、セマンティック検索の対象をウエブ（非構造化データ）にまで広げる可能性も十分残されている（詳細は後述の「6　非構造化データの解析にはAIが有効」を参照）。その場合、いわゆる「SEO（Search Engine Optimization）」と呼ばれる、検索エンジンに最適化したホーム・ページの記述方法も大きな影響を受けると見られる。つまり次世代ウエブ標準言語「HTML5」に用意された、セマンティック関連のタグ（要素）を最大限に活用した記述方法へと大幅にシフトすることになるだろう。

4　ソーシャルか、それともAIか？

それにしてもGoogleは、なぜ敢えて今、検索アルゴリズムの大幅改訂に取り組んでいるのか？それは同社の2大ライバルであるFacebookとAppleの動向を睨んでのことだ。この2社は今、Googleの従来型キーワード検索と真っ向から対立する「新たな情報へのアクセス方法」を世に問うている。

まずFacebookの方だが、彼らはいわゆる「ソーシャル・グラフ」と呼ばれる仮想コミュニティ内の人間関係を利用した情報入手方法を提起している。そこでは我々ユーザーはキーワードで機械的に情報検索する代わりに、「こんなことを知りたいんだけど、誰か知ってる？」といった形で問を投げかける。するとコミュニティの友人・知人が「それなら、これがいいよ。こんなこともあるよ。ここに行けばあるよ」といった形で答えを返してくれる。

あるいは「ニュース・フィード」のような形で、こちらが敢えて質問しなくても、自然に友達の情報が飛び込んで来る。これらの情報入手方法は、「コミュニティの構成要員の方が、検索エンジンよりも、自分と共通の関心、趣味、嗜好に合った答えを返してくれる」という基本思想に基づいている。

一方、Appleはこうしたソーシャル的アプローチとは対照的な手法を採用している。昨年発売のiPhone 4Sに標準搭載した「Siri」によって、いわゆる「AI（人口知能）」を使った新たな情報処理の在り方を提起したのだ。Siriは単なる情報検索の手段というより、もっと総合的な音声アシスタント機能の一種とされる。たとえばユーザーが「検索」、「メール」、「スケジュール管理」など様々な命令を音声で下すと、Siriは女性に模した合成音声でこれに応じ、まるで召使かロボットのように要求された仕事をこなしてくれる。

このSiriを実現するために使われている技術は、大きく次の3つに整理される。1つ目は、ユーザーの音声を認識する「音声認識機能」、2つ目は音声認識された結果として生成される言葉（命令）を理解する「自然言語処理」、そして3つ目はその命令に答えるために様々なウエブ・サービスへと仕事を割り当てる機能である。ちなみにSiriが行う情報検索の部分は、前述のWolframAlphaに大きく依存している。現在、Wolfram Alphaがこなしている仕事全体の25%は、Siriからクエリー（データベースへの問い合わせ）だ。

こうした中、GoogleはまずFacebookの脅威に対しては、「Google+」と呼ばれる自身のソーシャル・メディアを立ち上げることによって対抗しようとしている。しかしGoogleが本来、得意とするのは、ソーシャルな人間関係よりも、むしろソフトウエアのアルゴリズムをフル活用したAI的アプローチだ。そこで同社の屋台骨とも言える検索エンジンの部分にはAI的アプローチを導入し、この分野で先行したSiriに対抗しようとしているのだ（自然言語処理に基づくセマンティック検索は、AI的アプローチの典型である）。

5　セマンティック検索が狙うのは「ビッグ・データ」

進化する検索エンジンの背景にはまた、最近話題の「ビッグ・データ」が存在する。ビッグ・データの厳密な定義は存在しないが、印象としては現代社会に存在する、ほぼ全てのデータを網羅している感がある。

たとえば「企業が保有する大量の顧客情報や売上げデータ」、「政府機関や公共団体が保有する各種の統計情報やデータベース」、最近では「ブログやSNS、Twitterなどソーシャル・メディアに投稿される無数のコメントや写真」、さらに「各種センサーを搭載した様々な機器や装置から、時々刻々と上がってくる計測データ」など。要するに、これら諸々のデータを全てひっくるめて「ビッグ・データ」と呼んでいる。

そこではデータの分析（解析）が重要な意味を持つ。つまり大量のデータ自体に価値があるというより、それをいかに上手く解析して、企業や生活者にとっての新たな価値を創出するか。ここがビッグ・データの主眼となっている。

しかし、そういう事なら企業は昔からやっているのではないか――そんな感想を持たれる向きも多いと思う。たとえば「データ・マイニング」など、大量のデータを解析することで新たな収益機会を生み出す試みは以前から存在する。それと今回の「ビッグ・データ」とでは、何が違うというのだろう？

1つの答えとして、ビッグ・データの場合、たとえば「検索キーワードの地域的分布を解析することにより、伝染病の流行しそうな地域を予知する」、あるいは「政治的な動向や文化のトレンドを読む」など、多方面への応用が期待されている。この辺りが恐らく、これまでとは異なる点である。つまり単なるビジネスの領域を越えて、社会、政治、文化など多方面に渡ってデータ主導の科学的なアプローチを適用すること。これが最近の「ビッグ・データ」ブームのポイントだ。

以前との、もう1つの違いは、分析の対象となるデータの質的な拡大である。いわゆる「ビッグ・データ」は2種類に大別される。それは前述の「構造化データ（structured data）」と「非構造化データ（unstructured data）」である。もう一度繰り返すと、前者は、たとえば企業の情報システムに格納された「リレーショナル・データベース」など、体系的に組織化されたデータを指している。一方、後者はたとえばブログやSNSなどソーシャル・メディアに時々刻々とアップされていく、無数の書き込みや写真、動画など、多彩で無秩序な情報ストリームを指す。

そして現代が現代たる所以は、後者の「非構造化データ」が過去とは比較にならないペースで爆発的に増加していることにある。たとえば私たちが普段、家や職場、学校から呟くTwitterの書き込み、あるいはFacebook上で四六時中発するコメントや写真、そして「いいね」ボタンに代表される個人の嗜好情報など。これらは過去にはデータとして記録されてこなかったものであり、今だからこそIT企業のサーバー

にデータとして長期に渡って蓄積される。つまり、このような人間の心、情動などの非構造化データこそが、現代を象徴する「ビッグ・データ」の肝なのである。

この点から見て、今年5月のFacebook上場に伴う喧噪は、単にFacebookという一企業に寄せられた期待というより、もっと大きな「ビッグ・データ」という社会現象に対する世間の期待値を反映していると言えるだろう。そして上場後の同社株価が低迷したことは、このビッグ・データに寄せられた大きな期待が、ひょっとしたら期待外れに終わるという懸念も暗示している。つまりビッグ・データから経済的価値を引き出すのは、そう容易なことではないのだ。

6　ビッグ・データの解析にはAIが有効

ビッグ・データ、特にその主要部分を構成する非構造化データは、それを解析するために高度な技術を必要とする。たとえばFacebookに投稿され、ニュース・フィード上を流れる無数のコメントを解析するには、「自然言語処理」の技術が必要とされる。同じく、そこに投稿される写真や音楽、動画などを分析するには、「画像認識」や「音声認識」などの技術。さらに、これらの結果を総合して、そこから、ある種のパターンを抽出するには「機械学習」も組み合わせる必要がある。これらの技術はいずれもAIの代表格であり、1950年代から地道な研究開発が継続され、ここに来て、ようやくその成果が目に見える形で出始めたところだ。

AppleやGoogleが今、競って研究開発を進めているのが、このAI技術を使って「非構造化データの海」とも言えるウエブの世界から「セマンティクス（意味）」を引き出す技術である。従来の検索エンジンのような情報収集の仕方では、キーワードやリンクを機械的に分析して、そこから統計的な手法で、より適切と見られる情報をユーザーに提供しているに過ぎなかった。

が、このやり方では、膨大な非構造化データを解析して、そこから迅速で実践的な意思決定をする上での限界が見えてきた。むしろ検索エンジンやブラウザなど、いわゆる「エージェント」と呼ばれるソフトウエアが、ソーシャル・メディアなどウエブ情報の「意味」を理解した上で、それを解析することが必要となってきた。Googleのセマンティック検索は、まさにそこへと向かう第一歩と見ることができるだろう。

【執筆者プロフィール】

氏 名：小林 雅一 （こばやし まさかず）

所 属：KDDI総研

専門：メディア・IT・コンテンツ産業の調査研究

経 歴：東京大学大学院理学系研究科を終了後、雑誌記者などを経てアメリカに留学。ボストン大学でマスコミ論を専攻し、ニューヨークで新聞社勤務。慶應義塾大学メディア・コミュニケーション研究所などで教鞭をとった後、現職。

主な著書：

『日本企業復活へのHTML5戦略』（光文社）

『スマートフォンのすすめ―手のひらのクラウドで未来を生きる』（ぱる出版）

『ウェブ進化 最終形 「HTML5」が世界を変える』（朝日新書）

『モバイル・コンピューティング』（PHP研究所）

『社員監視時代』（光文社ペーパーバックス）

『欧米メディア・知日派の日本論』（光文社ペーパーバックス）

ほか多数。