解 説 1

ビッグデータ処理の展望

ー 変ぼうするデータ分析技術の動向 ー

ビッグデータというキーワードが近年 IT 業界を席巻しています。情報量の増加という現象そのものは情報処理の黎明期から連綿と続いており、特にインターネットの普及以降は、Web ページやソーシャル・メディア、画像・動画データ、センサー・データなどの非構造化データが驚異的に増加し、情報爆発と呼ばれるように広く認知されています。では、なぜ今ビッグデータが注目されているのでしょうか? その主な理由は、ビッグデータが単純なデータ量の増加を意味しているのではなく、データの多様性や発生の早さといった複数の属性を持ち、その取得や分析の方法によってこれまでになかった革新的なサービスやビジネス・モデルが実現可能だと理解されてきたためでしょう。

Apache Hadoop に代表されるオープンソース・ソフトウェアや、IBM をはじめとする各社から提供される製品・サービスの充実とともに、このようなビッグデータ処理の事例もさまざまな分野で見かけられるようになってきました。本解説記事では、現在もダイナミックに進化を続けるビッグデータ処理のエッセンスをご紹介し、どのような形でビッグデータの価値を導き出せるかについて、特に分析技術を中心にご説明いたします。

Article 1

Perspectives on Big Data Processing
- Changing Trends in Data Analysis Technology -

The “Big data” concept has recently been dominating the IT industry. Readers may wonder why it has gained such attention, since explosive growth in the amount of information has been consistently observed in this industry since the early days of information processing. In particular, the Internet has demonstrated extraordinary growth in “unstructured information” such as Web pages (text), social media, images, movies, and sensor data over the last decade. The question is Then why is there a buzz about Big data now. The main reason could be that the term “Big data” no longer means just the exponential growth of data, but indicates a variety of data types and the speed that data is being generated. We are beginning to understand that novel techniques for data collection and analysis could lead us to the innovative services and business models.

There has being a growing number of use cases of Big data in many areas and industries, as there are open source software such as Apache Hadoop as well as commercial products from IBM and other vendors that specifically address the processing of Big data. In this paper, we introduce the fundamentals of the processing of Big Data and how it can be used—primarily through advanced analytics—to drive value from it.

❶ ビッグデータ処理とは?

「ビッグデータ」の決定版といえる定義はまだ存在しませんが、直接的には年々指数関数的に増加するデータが、従来のストレージやデータベース・システムによって管理できる限界を超えつつあるという問題が顕在化したことが、この用語が誕生する背景となりました。McKinsey Global Institute（以下、MGI）が 2011 年 5 月に発行したレポート［1］では、「技術的進歩により扱うことのできるビッグデータの物理的なサイズは年々増加するため、典型的なデータベース・システムで収集・記憶・管理・分析することが困難なサイズのデータをビッグデータと呼ぶ」という相対的な定義をしていますが、現時点では数十テラバイトから数ペタバイトのデータがビッグデータとして扱われているようです。ガートナー社は、ビッグデータを**容量**（volume）、**種類**（variety）、**スピード**（velocity）の 3 種類の属性によって特徴付けており［2］、IBM でもこの定義を参照しています［3］。容量については前述の通りですが、種類にはデータベースに適合した構造化データ以外に、テキスト、音声、ビデオ、クリック・ストリーム、ログファイルなどのさまざまな種類の非構造化データが含まれます。また、スピードとは、データが発生する早さと、そのようなデータを処理しなければならないミッション・クリティカルな速度要求の両面を意味します。

技術的にこのようなビッグデータ処理が広く一般に意識されるようになった契機は、2006 年に発表された Google 社の BigTable に関する論文［4］でした。同社のインターネット検索を含む多数のサービスを技術的に支えているのが BigTable と呼ばれる、大規模かつ高速なデータベース・システムです。この後 KVS（Key-Value Store）と呼ばれるデータの格納・アクセス手法と、MapReduce という並列計算手法が大きな注目を集

めるようになりました。BigTableを実装する基盤となったファイル・システムやMapReduceのアイデアに触発された米Yahoo!社のダグ・カッティング氏がJavaでこのような分散処理基盤を実装し、その後Apacheソフトウェア財団においてオープンソース化されたものがHadoopです。世界最大のSNS（Social Network Service）会社であり、世界中で8億人を超えるといわれる会員を擁するFacebook社は、コンテンツの共有からアクセス・ログの解析、メッセージ機能などの開発といったビッグデータ処理をHadoop上に実装しています。2010年には日本でもMapReduceやHadoopに関する詳細な実証実験報告書［5］が公開されており、これらの基盤技術が本格的に普及する予兆を示しています。

IBMではビッグデータがもたらすビジネス変革の可能性に早くから注目してきました。ストリーム型データの処理については、2003年に研究プロジェクトを開始し、リサーチ部門の技術予測であるGlobal Technology Outlook（GTO）では2006年に「イベント・ドリブン・ワールド」というトピックで取り上げています。このストリーム型データ処理は、2010年3月にIBM InfoSphere Streams（本誌46ページ以下：解説②参照）として製品化されました。また、2010年にIBMが開催したInformation On Demand（IOD）コンファレンスでは、ビッグデータ処理のための新たな製品IBM InfoSphere BigInsights（本誌46ページ以下：解説②参照）が発表され、2011年のGTOでは「ペタ・スケールのアナリティクスとアプライアンス」というトピックでビッグデータの技術を展望しています。

ビッグデータ処理では、その膨大なデータ処理量のみが注目されやすいのですが、刻々と変化する情報や多様な情報から総合的な分析を行う手法が、知見獲得や意思決定支援の新たなソリューション（本誌36ページ以下：インタビュー④参照）として実現されつつあることを理解するべきでしょう。

以下では、主に数値データを深く分析することで可能になった数理科学的なビッグデータ解析手法を詳しく紹介し、さらに非構造化データを代表するテキストの分析についてご説明します。最後に、今後予想されるビッグデータ処理の展開を簡単にご紹介します。

❷ ビッグデータの分析

大規模データの分析を考える際の本質的な問いは、「本当にビッグデータを『Big』なままで取り扱う必要があるのだろうか」というものです。例えば、気象予測を行うために地表に風力計を設置する状況を考えてみます。この場合、風力計の個数と時間当たりのデータ取得回数の積に比例してデータ量は増えますが、東京の明日の天気を知りたい、というような大ざっぱな問いに対しては、各市町村に1～2個の風力計を置いて、せいぜい10分おきにデータを取得すれば十分でしょう。分析に必要なデータの分量は、「何をしたいか」「何が得られるか」に依存します。

それでは近年のビッグデータをめぐる活発な動きの裏には、どのような分析手段があるのでしょうか。それを考える上で重要なのが「ロング・テール」という概念です［6］。これは、従来型の、選択と集中による大量販売に対するいわばアンチテーゼとして主張されているもので、その最も華々しい成功例がインターネット書籍販売です。ここでは購買推薦の仕組みが本質的な役割を果たしています。購買推薦とは、ある任意の顧客に対して、その人が最も興味を持ちそうな商品のリストを提示する仕組みです。これは一見単純な作業に見えますが、裏で動いているのは、**協調フィルタリング**［7］と総称される洗練された機械学習のアルゴリズムです。協調フィルタリングでは、（購買者の数）×（商品の数）という巨大な表形式のデータを扱う必要があります。購買者数も商品数も100万のオーダーになり得るので、場合によっては兆のオーダーの要素を持つ巨大なデータとなります。そのような巨大データは保持するだけでも大変ですので、商品や顧客を間引いたりして、データのサイズを削りたくなるのですが、それは許されません。まさにそれがロング・テールたるゆえんなのです。たとえまれにしか売れない商品であっても、超多品種販売というモデルにおいては、あらゆる商品に対するすべての顧客データが、購買推薦のための重要なヒントを与えるものとして不可欠です。このような状況が、強いビジネス的動機付けを得て出現したというのが、ビッグデータの分析においてまずは重要なポイントです。

実はロング・テールの概念は、購買推薦のような新しいビジネス以外でもしばしば重要となります。例えばネットワーク監視というタスクを考えてみます。この場合、興味があるのは正常時の振る舞いというよりはむしろまれに生ずる何かの異常な状況で、全体の分布のすそ（tail）の部分に存在する現象が重要です。実際、データ・サイズを減らすために監視対象のパケットを間引くことは、アタックの検知性能に深刻な影響を及ぼすことが知られています［8］。同様の状況は、製造、運輸など、広範

な産業領域で得られる物理センサーからのデータにおいても生じます。次節で解説するテキスト・データと異なり、一般にセンサー・データそれ自体の可読性は低いので、そこから有用な知見を引き出すためには、高度な分析技術が必要となります［9］。**センサー・データ解析**、特にその異常検知技術は、ビッグデータに関するひとつの主戦場になっていくと考えられます。

最近のビッグデータに関係するもうひとつのトピックとして、いわゆる位置ベースのサービス（location-based services）が注目を集めています。これは、どういう人がどこにいるかという位置情報を用いて、例えばタイムリーに広告を展開しようというサービスです。そのような**時空間データ解析**は、店舗設計、商圏の設計などに有用と考えられています。この場合、通常、GPSトラッキング・データのような、位置情報を表すセンサー・データを分析することが必要になりますが、位置情報は時々刻々得られるため、地図情報も含め一般的にそのデータ・サイズは巨大になります。巨大な時空間データから意味ある知見を引き出す技術は発展途上というべきですが、これもまた最近のビジネス・モデルが要請する新しいビッグデータの問題といえるでしょう。最近の興味深い研究例としては、購買推薦の仕組みを時空間データに拡張した研究があります［10］。

以上、ビジネス・モデル、分析手法、データという三者の相互作用が、ビッグデータを「Big」なままで取り扱うことを必須とする新しい状況を生み出していることを見てきました。当然ながらこのことは、データ分析側にも多大なる反作用を及ぼしています。最近の顕著なトレンドは、機械学習と大規模並列計算環境の融合です［11］［12］。すなわち、並列計算環境を前提として、そこで使いやすいようにアルゴリズムの側を再構築する動きが急速に進展しています。実際、上記の文献では、ビッグデータに関する多くの計算タスクが、大規模並列計算環境の利用により高速化されたことが報告されています。

しかしながら、そこで明らかになりつつあることは、難しい課題を並列化により高速化することは簡単ではないという事実です。結局問題は、以前から連綿と研究されてきた大規模数値計算（HPC: High-performance computing）と同じ地点に逢着（ほうちゃく）してしまった感があります。例えば、前述した協調フィルタリングにおいては、特異値分解という演算を行うものがあります。実はこれは量子力学の基礎方程式であるシュレーディンガー方程式を解く際にも現れます。物理学におけるHPCの数十年にわたる苦闘の歴史から考えれば、それを並列化により劇的に高速化するのはかなり難しいでしょう。

向こう数年間で、情報検索におけるPageRankアルゴリズムや、購買推薦における強調フィルタリングのような、ビッグデータに対するいわば「キラー・アルゴリズム」が幾つも開発されることになるでしょう。しかしながらわたしたちは、ビッグデータの解析に魔法はないということを認識すべきかもしれません。ビッグデータを保持し、並列計算を可能にするインフラができればすべてが解決するわけではありません。ビッグデータを活用するためには、今後ますます、分析手法それ自体の新規開発能力が重要になっていくことでしょう。

❸ 大規模テキスト・データの分析

大規模テキスト・データの分析は、2000年代前半にコールセンターの会話ログ分析およびレポート作成支援で最初のブームを迎えました。当時のコールの会話テキストは年間数十万件から数百万件にも達していたため、よくある質問の抽出や、製品の反響や初期不良の早期検知を人手中心で作業することは不可能になりつつあり、これらのデータの大半が分析に利用されずに死蔵されていました。テキスト分析技術はこのような状況を一変させ、CRM（Customer Relationship Management：顧客関係管理）分野に顧客の声分析という手法を確立させました。売上・販売予測などのダッシュボード情報にテキスト分析結果を関連付けることで、消費者の不満、満足や要望に関する意見を理解し、数値データが示唆するマクロ的な現象の要因や影響についての深い洞察が可能になったのです。このような手法は、金融業界における苦情分析や自動車業界における不具合分析などにも発展しています。

現在ではテキスト分析の対象がTwitterやFacebookなどの**ソーシャル・メディア**と呼ばれるコミュニティー中心の情報共有・交換のテキスト・データに移りつつあります。消費者の商品購買を含む重要な意思決定にソーシャル・メディアの口コミが大きな影響を与えることが分かってきたためです。2011年10月に発表された株式会社電通の調査では、インターネット利用者の4割がその購買行動にソーシャル・メディアからの影響を受けていたことが示されています。情報漏えい、デマや風説の流布も含めると、企業のマーケティング、リスク管理、情報のガバナンス、品質保証、商品企画などの広範な

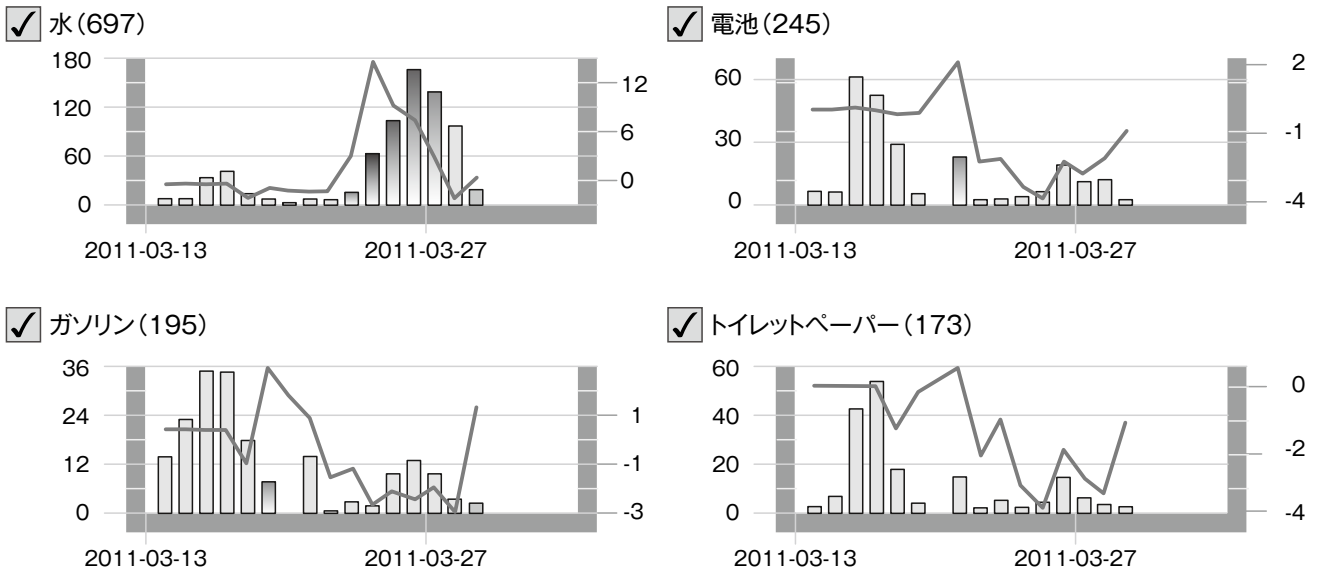

図1. 震災後の不足物品の分析

業務にソーシャル・メディアを強く意識することが不可避となるでしょう。

これらのソーシャル・メディアは本格的なテキストのビッグデータといえるでしょう。例えば、1日当たりに世界中で生成されるTwitterでのtweet（投稿）の総数は2011年6月の時点で2億件を突破したと報告されています。公開されている検索API（Application Programming Interface）を利用して、収集対象を特定のキーワードを含むtweetのみに制限しても1日当たり数十万件のtweetを取得できると考えられますので、件数だけでいえば中規模のコールセンターで1年間に蓄積されるメッセージと同数のメッセージが1日で集まることになります。

図1は、このようなソーシャル・メディアから、2011年3月の東日本大震災直後の数週間にわたり「～が買えない」や「～が売り切れ」といった物資の不足を訴える表現を抽出し、その不足物資（ここでは水［697件］、電池［245件］、ガソリン［195件］、トイレットペーパー［173件］の4品目）の出現頻度を時系列で表示したものです。物資ごとに固有の窮乏パターンを示していることが分かります。この手法は医薬品不足の情報を収集するプロジェクトにも利用されました。同様の手法は、感動を表す表現を集めたスマートフォン向けアプリケーションでも使われるようになっており、ソーシャル・メディアを利用したサービスの多様性を感じることができます。

ソーシャル・メディアのテキスト分析については、ProVISION 70号48ページ以下の解説②［13］ですでにご紹介していますが、従来のコールセンター向けの顧客の声分析と大きく異なる点として、以下の4つを挙げることができます。

- 特定の話題が、人やコミュニティーのネットワークを通して急速に広まる可能性がある。
- 話題ごとに、発信者とそのフォロワーのネットワークに依存して異なる影響度を持つ。
- モバイル機器からの投稿を反映して、投稿者や話題と位置情報に強い相関を示すものがある。
- 企業が発信者となってキャンペーンやマーケティング活動をした反響を知ることができる。

従って、単にテキスト情報のコンテンツやデータ量だけではなく、時間空間的属性、人やコミュニティーのネットワーク構造、メディアの持つ双方向性などの複雑な属性を考慮した分析が必要とされることが分かります。ほかにも利用者の行動特性（コンテンツへのアクセス、発信、広告のクリックなど）とテキスト分析とを組み合わせた手法も提案されつつあり、今後はeコマースにおける新たなユーザー体験やビジネスの創造が期待されています。ビッグデータにより、ライフログ的に個々の利用者の特性が細密化されれば、プライバシーへの十分な配慮は必要ですが、よりパーソナルあるいは状況にフィットした推薦やサービスが提供できるようになるでしょう。

またソーシャル・メディアの双方向性を利用したキャンペーンとその効果のリアルタイム分析も見逃せない傾向

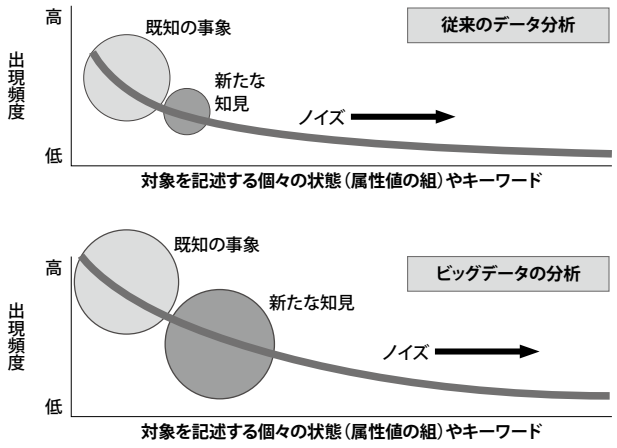

図2. 従来のデータ分析とビッグデータ分析の違い

です。企業が発信するメッセージがどのような消費者にどのように伝達されていくかをより深く把握する上で、デジタル・マーケティングはビッグデータ分析と強く結び付いていくでしょう。

❹ ビッグデータ処理の今後の展開

これまでにご紹介したビッグデータ処理の要点を直観的にまとめると**図 2** のようになるでしょう。従来のデータ分析では対象とする世界を記述する状態（属性の値の組み合わせ）やキーワードのサンプル数が十分ではなく、高頻度で出現する既知の状態を除けば、新たな傾向や異常状態などの重要な知見は偶然観察されたノイズとなる状態に埋没していました。このため新たな知見の獲得は極めて困難な作業だったのです。ビッグデータの分析では、その豊富なサンプル数から、対象世界の有益な知見がノイズと区別できる可能性が高まり、情報の種類や発生速度などの特徴と合わせて知見獲得の機会が増えると考えられます。特に大規模で多様な対象を分析するときにこの差が決定的なものとなり、個別の顧客へのサービスや、システムの異常検知能力を差別化することができるでしょう。

EMC 社の調査［14］によると、ビッグデータをビジネスに活用している企業はまだ全体の 1/3 にとどまるとのことです。同調査は、欧米や中国の企業におけるデータ分析の専門家約 500 名からの回答に基づいており、データ分析のスキルを持つ人材の不足が深刻な状況であることを印象付けています。前述した MGI の報告［1］でも、米国で 2018 年までに 14 ～ 18 万人程度の人材不足が起こると予想しています。これまでもビジネス・インテリジェンス（以下、BI）を分析の専門家のみでなく、業務担当者が利用しやすいように使いやすさ（consumability）を向上させるように努力はなされてきましたが、今後ビッグデータを活用するためには、BI ツールにより強力なデータの集約・可視化手法や業務向けの利用パターン［2］を提供する、業務アプリケーションに応じてビッグデータ分析を利用可能にする、といったことが求められていくでしょう。

ビッグデータの収集、保存・管理、分析のための技術的な進展も加速されるでしょう。半導体ディスク（SSD）やメモリー上で動作するインメモリー・データベース、容易に追加・拡張が可能なアプライアンスといったコンポーネントを駆使したビッグデータ処理基盤が続々と市場に投入されることも予想されます。

このようなビッグデータを活用した事例は、今後製造業、情報産業、エネルギーと医療などの分野で先行すると予測されています［1］。このほかにも、IBM のスマートな都市を実現する IOC（Intelligent Operation Center）［15］のように、ビッグデータの集積と分析によって管制塔のように効果的に都市の行政機能を構築できることを示唆する事例があります。コンシューマー分野では、クラウド基盤でビッグデータ処理を活用した独創的なインターネット上のサービスが続々と登場するようになるでしょう。スマートフォンの普及とそれに続くモノのインターネット（Internet of Things）の時代に備えて、ビッグデータの活用を戦略的に構想する時期が到来したといえるでしょう。

［参考文献］

［1］ James Manyika, et al.: “Big data: The next frontier for innovation, competition, and productivity,” McKinsey Global Institute, http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation（2011-5）.

［2］ Gartner: “Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data,” http://www.gartner.com/it/page.jsp?id=1731916（2011-6）.

［3］ IBM：“ビッグ・データとは?,” http://www.ibm.com/software/jp/data/bigdata/（2011）.

［4］ Fay Chang, et al.: “Bigtable: A Distributed Storage System for Structured Data,” Proc. OSDI'06: Seventh Symposium on Operating System Design and Implementation, http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//archive/

bigtable-osdi06.pdf（2006-11）.
[5] 株式会社エヌ・ティ・ティ・データ：経済産業省平成 21 年度産学連携ソフトウェア工学実践事業事業成果報告書，http://www.meti.go.jp/policy/mono_info_service/joho/downloadfiles/2010software_research/clou_dist_software.pdf（2010-3）.
[6] Chris Anderson, The Long Tail: Why the Future of Business is Selling Less of More, Hyperion,（2006）.
[7] F. Cacheda, et al.: Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems, ACM Transactions on the Web, Vol. 5, No. 1, pp.2:1-2:33,（2011）.
[8] J. Mai, et al.: Is sampled data sufficient for anomaly detection?, Proceedings of the 6th ACM SIGCOMM conference on Internet measurement (IMC '06), pp. 165--176,（2006）.
[9] 井手剛：“スパース構造学習によるセンサー・データの変化点検出と異常解析，” PROVISOIN No.65, http://www.ibm.com/ibm/jp/provision/no65/pdf/65_paper1.pdf（2010-Spring）.
[10] R. Raymond, et al.: Location Recommendation based on Location History and Spatio-Temporal Correlations for an On-Demand Bus System, Proceedings of 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems,（2011）.
[11] A. Ghoting, et al.: SystemML: Declarative Machine Learning on MapReduce. Proceedings of the 2011 IEEE 27th International Conference on Data Engineering (ICDE 2011), pp. 231-242,（2011）.
[12] A. Ghoting, et al.: “NIMBLE: A Toolkit for the Implementation of Parallel Data Mining and Machine Learning Algorithms on Mapreduce.” In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2011), pp. 334-342,（2011）.
[13] 米持幸寿：“震災時ソーシャル・ネットワークの効果と脅威 ―評判・風評分析の重要性 ―，” PROVISOIN, No.70, http://www-.ibm.com/ibm/jp/provision/no70/article2.html（2011-Summer）.
[14] EMC: “New Global Study: Only One-Third of Companies Making Effective Use of Data,” EMC Press Release, http://www.emc.com/about/news/press/2011/20111205-02.htm（2011-12）.
[15] 小林真：“Intelligent Operations Center ― 協調的問題解決を支援するシステム基盤 ―，” PROVISOIN, No.71, http://www.ibm.com/ibm/jp/provision/no71/pdf/71_article1.pdf（2011-Fall）.

日本アイ・ビー・エム株式会社
東京基礎研究所　(IBM Research – Tokyo)
技術理事 (Distinguished Engineer)

武田 浩一　**Koichi Takeda**

[プロフィール]
1983 年、日本 IBM 入社。以後、東京基礎研究所において自然言語処理やテキストマイニングの研究開発に従事。英日機械翻訳システムやテキストマイニング・ツールの研究開発に貢献。2007 年 12 月より Watson プロジェクトに参加。現在は質問応答とテキストマイニングの統合や、電子カルテなど医療情報のマイニングといった新しいビジネス・インテリジェンスの実現に取り組んでいる。

日本アイ・ビー・エム株式会社
東京基礎研究所　(IBM Research – Tokyo)
数理科学担当 (Mgr. of Analytics & Optimization)

井手 剛　**Tsuyoshi Ide**

[プロフィール]
2000年にIBM東京基礎研究所に入所。液晶工学の研究に従事。その後、データ・マイニングの研究に転じ、2004 年ごろから、センサー・データ解析のグループをリード。主に時系列データの異常解析技術に関する研究に従事。人工知能学会全国大会優秀賞（2004、2006 年）ほか受賞。博士（理学）。