■研究論文

コンテクストマイニング

—その対象とBI構築の方向性—

A Study of Methodologies and Tools for Context Mining to Build Effective BI (Business Intelligence) Systems

立教大学 佐々木 宏
Rikkyo University Hiroshi SASAKI

1. はじめに

データウェアハウスの概念が登場してから，すでに10年以上も経過している．その間テキスト・マイニング，Webマイニング，BI（Business Intelligence）など，実務への適用は確実に進展してきた．時代によらず共通するのは，データ・マイニングが取引の背後にある商品，顧客などの見えざる構造の視覚化を目指しているという点である．

データ・マイニングのために活用されるのは，これまでPOSデータなどトランザクションから収集された定量的データが中心であった．たとえば，『紙おむつとビールの関係』は，たしかにひとつのコンテクストを発見しているが，それが発掘できるのは多くのデータが集積された後で，スピードは遅い．そこで，より一層必要とされているのは，鮮度の高い「状況把握」のためのデータ活用である．

コンテクストには前後関係，文脈，背景，状況などの意味があり，情報はコンテンツであって，同時にコンテクストを形成していくところに特徴がある．単純なメールの交換から，eコマース，多対多のコミュニケーションに至るまで，サイバー社会には多種多様なコンテクストが保存されている．すると，データ・マイニングとはデジタルデータというコンテンツを活用して市場や社会のコンテクストを発見する手続きだ，といい換えることもできる．本稿で問題にするのは，そのようなコンテクストを発掘するマイニング手法と，それをベースにしたBIの構築原理である．

2. 従来のデータ・マイニングとコンテクストマイニング

2.1 データ・マイニングの限界

従来のデータ・マイニングでは，業務トランザクションの大量データを収集した後，統計などの分析手法を使ってその実体を明らかにし，意思決定に役立てようとする．しかし，そこにすでにコンテクストマイニングを困難にする要因が含まれていることに着目したい．

情報システムの設計プロセスに立ち返ってみると，まずシステム化のターゲット（戦略目標）を明らかにし，その範囲を規定して全体モデルの概念設計を行う．次に，情報流のフローとストック

に着目して，システムの構造とプロセスの詳細を定めるという手順で構築が進む．その際の代表的なアプローチとして次のようなものがある．

①プロセスオリエンテッドアプローチ：業務手続きの側面に着目して，これを完全に記述する．

②データオリエンテッドアプローチ：業務データの側面に着目して，これを完全に記述する．

③EA（Enterprise Architecture）：組織の構造と業務プロセスを整合させ，全体をくまなく表記する．

④UML：業務をオブジェクトの集積とみて，その振る舞いをすべて規定する．

これらの背後には，いずれもコンシステンシーとインテグリティが保証された美しいモデルが想定され，それを完全に記述できることを暗黙の了解としている．しかし，上記のプロセスで構築される情報システムは完全ではなく，すべての業務の流れの一部を切り取っているに過ぎない．このことは情報システムの実装・運用段階で露呈する．

①システムの不完全性

・業務仕様は不完全であり，例外処理が発生する

・技術（IT）は不完全であり，人間系の関与が不可欠である

・ビジネス環境のスピードは早く，システムの改変が追いつかない

②情報の不完全性

・契約は不完全（不完備）であり，顧客とのコミュニケーションは不可欠である

・トランザクションデータの集積に時間がかかり，その間，市場はすでに変化してしまっている

・業務上の情報すべてをITで管理し，完全に分析するのはほとんど不可能である

システムの「全体性」に注目すると，商取引は，取引関係（B2B, B2C, C2C）で閉じたシステムであり，それを取り巻くコミュニケーションは，もともと社会に開かれて（社会全体でのみ閉じて）いる．いくら法律や制度をつくって規制しようとしても，個人情報保護が完全にできないのは，コミュニケーションは社会にオープンであるため，それを企業内部でクローズ化することが困難だからである．顧客同士のコミュニティができて販売元企業が情報をコントロールできない事態が生じるのも同じ理由からであり，だからこそ，サイバー社会のコミュニティと共存するビジネスモデル（國領，1998）が必要になる．

通常の業務システム設計では，本来不可分なトランザクションとコミュニケーションの情報体系からトランザクションデータを切り出して，そのなかで閉じたシステムを構築しようとする．その際，システム内部の合理的な手続きに則って業務が進む場合には，業務効率を劇的に改善することができる．ところが，業務パッケージの仕様に合わせて業務フローを変更したが，うまく現場が適応できず，却って部門間のやりとりに手間がかかるようになってしまった，などの状況もしばしば起きる．この例は，ITベースの情報システムの不完全部分を補完するために，システム手続き外の人間系のすり合わせ（＝コミュニケーション）に頼らざるを得なくなり，全体効率性を損なう可能性があることを示している（図1）．

大量のトランザクションデータを利用した事後的マイニングの限界は，ここから明らかになる．冒頭の『紙おむつとビール』について再度取り上げると，マーケティング担当者が店員から「また，近所のおじさんが紙おむつを買いにきた．休日で奥さんに買い物を頼まれているのだな．必ず，ご褒美のビールを買っていく．」というメールを受け取れば，それがPOSデータの分析結果とほぼ同じ価値を有していることがわかるであろう．トランザクションデータから，「紙おむつ」と「ビール」のような顧客の購買行動のコンテクストを再構築するには相当の時間とデータの集積を必要とする．トランザクション同士の関係をみるために，会員カードなどを発行して（ID付きにして），独立したデータを結びつけることもできるが，それでもなお，分析できるのはリピートの多い得意客のRFMなど，決まりきったコンテクストに限られる．トランザクションデータのマイニングでは，むしろ刻々生じている取引から瞬時にその変化や

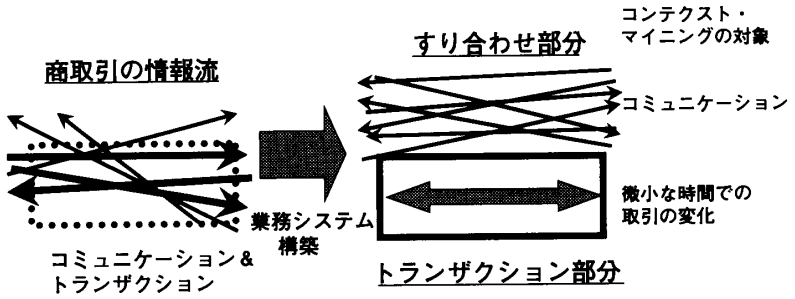

図1　コンテクストマイニングの対象

状況を把握し，すばやく意思決定に反映させるところに新しいニーズがある．微小な時間での取引コンテクストの把握，これは市場と製品の時間軸上の「すり合わせ」であり，コンテクストマイニングの対象の1つとなる．

2.2　コンテクストマイニングの方法

コミュニケーションのやりとりはロジカルなものではない．いわばスパゲッティ状態のコンテクストを構造化，可視化するための方法にはどのようなものがあるか．「コンテクスト」と名づけられた既存の概念や手法について，異なる領域から3つを簡単に取り上げたい．

まず，マーケティング関連の「ブランド戦略シナリオ」（阿久津・石田，2002）では，企業・顧客・社外関係者のコンテクストを構造化・内部共有化し，ブランド戦略として新たなコンテクストを創造するプロセスが論じられている．具体的ステップでは，コンテクストを探索，構造化，推敲して戦略シナリオの作るPLANのステップ，次に作り上げたコンテクストを内部共有化し，市場とのインタラクションを通じて刺激，協創するDOのステップ，最後にコンテクストを管理するSEEのステップが提示されている．コンテクスト構造化のツールとして指摘されているのは，テキスト・マイニングの利用である．

次に，現在のWeb環境からみるといささか古いが，IT関連として「コンテクスチュアル・デザイン」（Beyer et al., 1998）という方法論がある．そこでは，いかにして顧客主導（customer-centered）のITシステムを設計するかが述べられていて，コンテクスト探求のために，顧客（エンドユーザ）の現場に行き，どのような具体的データが必要かを把握すること，顧客とコラボレートして業務を理解すること（パートナーシップ），業務構造と可能な支援システムの構想（インタープリテーション），明確に焦点を絞ってインタビューを実施すること（フォーカス）の4原則が挙げられている．この方法は，顧客データを収集し，それをもとにシステムを設計するというもので，従来の情報システム設計の方法論を超えるものではない．また，顧客分析からアイデア創発に至る部分はプロジェクト・チームの，いわば人間系の活動に頼っていて，利用できるツールについては特に明示されていない．

さらに，ユビキタス社会の到来に関連して，人工知能の分野で活発に議論されているコンテクスト・アウェアネスと呼ばれる概念がある．コンテクスト・アウェアネスとは，IT技術を活用して生活空間上の状況認識を行い，的確な指示を出すような仕組みをいう．自宅のさまざまな場所にセンサーを設置しておき，ドアに近づくと自動でドアが開く，雨が降るという天気予報が出ているときに，外出しようとすると傘を持っていくよう指示が出る，などが一例である．

3. コンテクストマイニングの構築原理

ここまでの議論から，発掘すべきマイニングの対象を4つの次元に分類，整理する．コンテクストマイニングは，市場やユーザとの「すり合わせ」に重要な部分があり，さらにサイバー社会や実社会からのシグナル受信に新しい可能性が拓けてきたということができる．

3.1 コンテクストマイニングの次元

1つ目の次元は，従来と同じ商取引そのもののコンテクストである．企業活動はすべてプロセスであり，市場に投入した製品に関わるコンテクストは刻々変化している．商品の探索と購入意思決定の結果，取引が成立し，IT上のトランザクションとして結実する．取引状況と市場の変化を瞬時に把握し，それを意思決定に役立たせていくところにコンテクストマイニングの対象がある．

2つ目の次元は，商取引に付随するコミュニケーション部分のマイニングである．商取引にとって，コミュニケーションは不可欠であり，業務担当者は定型化された受発注，請求支払いなどのトランザクションの前後，すなわち商談やアフターサービスなどの場面で非定型的なやりとりを行っている．それ以外でも，コールセンターの対応，FAQのアクセス履歴など，情報システムの構築プロセスで排除されてしまったコミュニケーション部分に発掘すべきコンテクストが存在している．

3つ目の次元は，サイバー社会のコンテクストである．広義にオープンソースを捉えるなら，種々のコミュニケーションのコンテクストが，だれもがアクセスできるコンテンツ（デジタル情報）としてサイバー社会にくまなく遍在している状況だと考えることができる．通常，オープンソースということばで想定されるのは，Linuxなどのソフトウェアのソースコードのオープン化であるが，これは広義のオープンソースの特殊な形態で，ソフトウェアという特定のコンテンツ創造の目的に対し，その理念や協働のしくみなどのコンテクストが社会で共有化され，機能する例を示している．たとえば，市場に製品を投入すると，ただちに多くの風評がWeb上に流れ始め，それが自己組織化しさまざまなコンテクストを形成していく．たったひとつの書き込みで商品ブランドに傷がつくこともある．逆に，ネット上の口コミにより急激な普及がもたらされたりもする．それらが製品ライフサイクルに与える影響は少なくなく，ここに事業戦略上極めて重要なヒントが隠されている．

4つ目の次元は，実社会の状況把握としてのコンテクストである．コンテクスト・アウェアネスの技術は，生活空間からシグナルを得て，コンテクストを解釈する点でリアルな世界とバーチャルな世界の接点部分に位置づけられる．たとえば，RFID（Radio Frequency Identification）タグを人間の所持品，カート，商品に取り付けて顧客の店舗内の動きを把握したり，レジの自動化を行うなど，大手小売業の取り組みが本格化しつつある．

3.2 コンテクストマイニングのためのデータ構造

コンテクストマイニングのためのデータはどのような構造に集約できるか．その際，これまでのマイニング技法が，可能な限り活用できることが望ましい．上記4つの次元に対応させて，①トランザクションデータ，②商取引に関連するコミュニケーションデータ，③サイバー社会のオープンソース，④実社会からの検知データに分けて考えてみる（図2）．

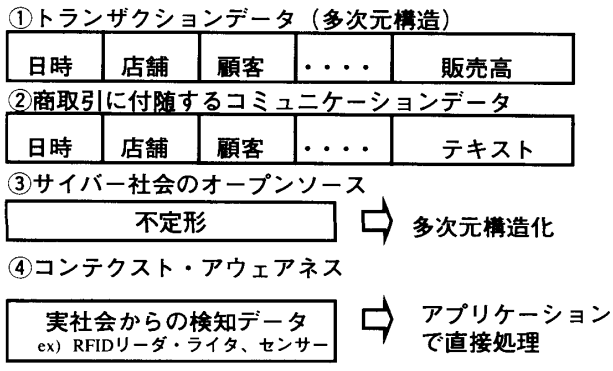

図2　コンテクストマイニングと4つのデータタイプ

まず，①のトランザクションデータは原データを多次元構造化させ，これをクレンジングした後，データウェアハウスを経てデータマートへ格納するのが一般的である．次に，②ではトランザクションデータの販売額に該当するところがテキストデータであるところに違いがあるものの，同じ多次元構造化が可能で，従来のドリルダウン，スライス＆ダイスなどの手法を適用できる．最も大きく異なるのはテキストデータ部分の分析である．

a) コンテンツ内コンテクスト：顧客からの意見など，一度のデータ収集機会で獲得できるコミュニケーションデータは，それ自体にコンテクストを内包している．文章内（1コンテンツ）のストリームを適切に分析する仕組みが必要である．

b) コンテンツ間コンテクスト：Web上のコミュニケーション・プロセスに何らかの仕掛けやルールがあって，コンテンツ同士が結合されていくものがある．Web訪問履歴，ブログでの閲覧者からのコメントやトラックバックの仕掛け，FAQへのアクセスなどは，一定のルールに従ってそのプロセスが保存されていく．

③が対象とするのはサイバー空間のあらゆるコミュニケーションである．日々の生活で，あることを知りたいと思えば，適当な掲示板で質問して直ちに回答を得ることができるし，何かを買いたいと思えばクリックひとつで物品を購入することもできる．サイバー社会に存在している膨大なデジタルコンテンツは，常に膨張を続ける人類の断片化された知識のかたまり，あるいはひとつの意識体のようでもある．すると，マイニングとは，玉石混合のいわば巨大な意識体から芥子粒ほどの金塊を取り出す作業にも例えられよう．このとき，情報の価値は受け手によって異なるので，それが金塊であるのか単なる芥子粒であるのか，他人にはわからない．③のコンテンツは，テキスト，音声，画像，動画など一定の形式をもたないが，それにデータ取得日，URLなどの基本属性を付加することでやはり多次元化が可能である．

最後の④は，RFIDリーダ・ライタやセンサーなどを介して得られた実社会からの受信データである．このシグナルを直接アプリケーションに渡して処理をすることができる．

コンテクストマイニングの基本ステップを示せば図3のようになる．上記のデータ構造化に加え，語彙の統一やことばの解釈など「コンテンツ整備」には特に手間がかかる．このステップがマイニングの品質を決定するといっても過言ではない．一例を示すと，「この焼肉やばいね」というとき，肉の品質が悪いという意味か，正反対に「本当においしい」という意味かはまったく判定できない．このようなコンテンツを放置したままマイニングを実行すると，誤った判断を招きかねないことになる．

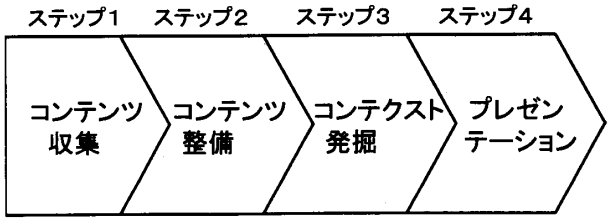

図3　コンテクストマイニングの基本ステップ

3.3　コンテクストマイニングのための技術

コンテクストマイニングのために必要となる技術のなかで，特にテキスト・マイニング関連に注目したい．第1はコンテンツ内コンテクストを分析するためのマイニング関連技術である．そこでは，文章の形態素解析，構文解析，シソーラス生成などのエンジン部分の品質が極めて重要である．形態素解析では，顧客評価などの基本的な分析において「よい」「良い」「いい」「いいかもしれない」「よくないかもしれない」「悪い」「良くない」「良くなくない」・・・など日本語独特の微妙な言い回しを正確に判定できなければならない．また，構文解析では，1文章内の単語同士の関連（主語述語の係り受け）をネットワーク構造化して可視化する，あるキーワードをもとにドリルダウンし

たり適宜原文を参照したりする，などの機能が求められる．

第2は，コンテンツ間コンテクストを分析する機能である．情報量を集約したり，相互関係を発見したりする方法として，同一文章内で出現した単語同士の類似性をもとに文章間の距離を定義してクラスタ分析し，結果を図示する機能や，コレスポンデンス分析を応用して単語同士，あるいは単語と属性との関係を二次元マップで視覚化する機能などがある．さらに，Web閲覧，FAQ，ブログなどそれぞれの仕掛けに対応して適切に分析する機能，さらにコールセンターに対応したアプリケーションや，多様なソースからの分析結果をポータルサイトのような形で表示させるアプリケーションの必要性も指摘できる．

ITサービス企業は，すでにこれらの要素技術を組み合わせたソリューションを展開している．各社が提供するテキストマイニング・ソフトウェアの機能は拮抗しつつあり，統計関連，オフィス関連，SIサービスなど，元来の事業とのシナジー部分で，それぞれ差別化をはかろうとしている．

第3は，各種タグを使って定点観測を行い，実社会からシグナルを獲得し，それを意思決定に生かすアイデアや技術である．ロジスティクスやマーケティングなど，この技術の応用できる分野は広く，実験段階を経ていよいよ普及期に入ってきている．先進的な事例として㈱NTTデータを取り上げると（白樫，2004），すでに社員がRFIDタグを付けてオフィス内での行動や状況を把握できるシステムを稼動させているほか，対外部サービスとしてユビキタス・サービス・プラットフォーム（RFIDプラットフォーム，CAM：コンテクスト・アウェアネス・マネジメント・プラットフォームなど）や多様なユビキタス・ソリューションを提供している．

4. まとめ

本稿では，従来型の業務トランザクションデータの事後的マイニングには限界があることをいい，商取引に付随するコミュニケーション部分に重要なコンテクストが潜んでいるところに着目した．ところが，コミュニケーションは本来社会にオープンであり，顧客はサイバー社会という深遠な世界や実社会というリアルな世界とつながっているため，より広範囲なマイニングが必要になる．

ここで，サイバー社会の日々膨張する巨大データの集合は，プラクティカルには加算不能と考えるのが適切であろう．ところが，アレフゼロ（加算可能な有理数の濃度）とアレフ1（加算不能な無理数の濃度）には本質的な断絶があり，それはかつて集合論を危機に陥れたほど大きい．すると，Webの検索エンジンは，その濃度の差を軽々と乗り越える魔法の杖のようにみえる．喩えていうなら，夜空という連続した空間から意味ある星座を見つけ出すようなイメージである．同様に，実社会も連続した時空間を構成しており，コンテクスト・アウェアネスの技術もまた，連続的時空間からシグナルを拾うという点では共通している．

空間軸とともに，考慮されるべきは時間軸の側面である．多くのデジタル機器をみればわかるように，商品ライフサイクルは極端に短く，普及曲線は美しいS字を描かず，著しく歪んだものになっている．製造業者と消費者との情報の非対称性は崩れ，消費者はいつどのようなスペックの製品がどのメーカーから販売されるかなどの情報を事前にキャッチする．ひとたびヒットが生まれると生産は追いつかず，店頭在庫は瞬く間に無くなってしまう．このような事実は，データを累積して回帰式を求め将来を予測する方法の有効性そのものに疑問を投げかけている．微小な時間でのトランザクションの変化から市場の動きを察知し，瞬時にマーケティング活動に反映させるコンテクストマイニングの業務は，次第に為替ディーリングのようなスタイルに近づいていくと考えられる．

最後に，コンテクストマイニングを実装するBI構築の方向性について要約する（図4）．これまでのデータ・マイニングでは，データ収集後，データ間の関係づけを再構築し，顧客行動や一般社会のコンテクストを発掘するところに力点が置かれた．ところが，本稿で述べたコンテクストマイニ

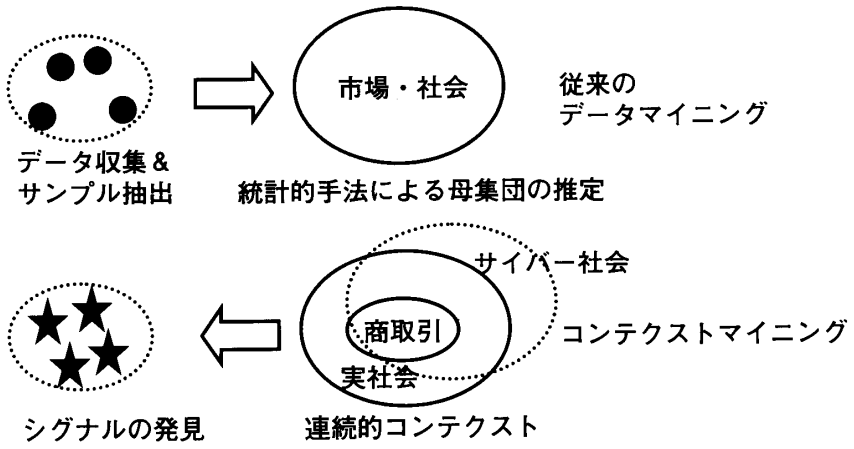

図4　コンテクストマイニング：過去の分析から現在の分析へ

ングでは，商取引を核にして，それを取り巻く一般社会，さらにそれとほぼ重なるように存在するサイバー社会，これらの連続的時空間のデータを抽出して，データを整え，即時にマイニングを実施する．マイニングの方向性は，過去の分析から現在の分析へというところに大きな変化がある(図4の矢印)．もちろん，従来の事後的データ・マイニングそのものを否定するわけではなく，その限界を補完し，両者を融合するBI構築が求められている．

謝　辞

本稿は，OA学会第50回大会の報告（佐々木，2005）を大幅に改訂したものである．報告の際に大阪市立大学大学院太田雅晴教授，神奈川工科大学田中宏和教授には，AI分野との関連などについて貴重なコメントをいただいた．記して感謝申し上げたい．

参考文献

阿久津聡・石田　茂（2002）『ブランド戦略シナリオ　コンテクスト・ブランディング』ダイヤモンド社.

國領二郎（1998）「顧客間インタラクションによる価値創造モデル」『ダイヤモンドハーバードビジネス』Oct-Nov, pp. 102–109.

佐々木宏（2005）「ソシオ・エコノミック統合型システム—オープンソース時代の情報システム設計に向けて—」『OA学会第50回全国大会予稿集』pp. 53–56.

白樫和明（2004）「コンテクスト・アウェアネスが実現するユビキタス・ネットワーク社会」『NTTデータ』http://www.nttdata.co.jp/ceatec/pdf/ct04_tn_02.pdf.（2005年7月21日）．この資料の他，NTTデータ取材の際の受取資料を参照.

Beyer, H. et al. (1998) *Contextual Design: Defining Customer-Centered Systems*, Morgan Kaufmann Publishers.