情報探索システム“ CoreExplorer ”を利用したテキストマイニング事例

Case of Text Mining using Information Retrieval System “ CoreExplorer ”

塚原　朋哉　Tsukahara Tomoya
佐藤　俊也　Sato Shunya
椿山　俊和　Tsubakiyama Toshikazu
高梨　勝敏　Takanashi Katsutoshi
井上　　悠　Inoue Haruka

テキストマイニングとは，大量の文書中のテキストから有用な情報を抽出する技術のことである。近年では，コールセンターに寄せられる顧客の声の分析などに用いられ始めている。

しかし，有用な情報が抽出されたかどうかの判断は分析の観点に拠って異なる。すでに様々なテキストマイニングのツールが存在するものの，対象データを様々な観点で簡単に分析できるシステムを求める声は多い。

本報告では，検索結果として文書のタイトルと文書を特徴付ける特徴語を自動的に出力する情報探索システム“ CoreExplorer ”を利用したテキストマイニング事例について報告する。

“ CoreExplorer ”は，文書の本文中の単語や属性を利用して分析対象を細かく指定したり，様々な観点で絞り込んだ分析を行うことが可能である。分析対象から抽出する情報は，特徴語同士の関連や文書同士の関連，特徴語と文書の関連の情報である。また，抽出する特徴語として文書の個所・属性等を選択することもできる。本システムを適用した２件の事例から，本システムが顧客要求にあった動作をしていることを確認した。

1　はじめに

本や雑誌などをはじめ，企業内部の文書など大量の文書が電子化され蓄えられており，今後もその量は増大する。蓄積文書数が増えるに従い，大量の文書中から，自分にとって価値のある情報を探し出すことが困難になり，また，社内の財産（蓄積されている文書）の活用も難しくなる。

アンケートの集計やコールセンターへの問い合わせ記録の分析では，属性や数値などのデータに対しては，多変量解析を用いた統計的な手法の分析システムが多くあるが，自由記述文などの自然言語に対する分析は，まだ人手に頼ることが多く，回答数が増大すると自由記述文に書かれた意見の見落としや，全体的な傾向が把握しづらいなど，調査結果を十分に活用できない。

これらの文書や自由記述文などのテキストから有用な情報を抽出したり，情報の傾向を確認する技術にテキストマイニングがある。テキストマイニングを行う分析システムでは，大量のテキストから情報を特徴付けるキーワード（以下特徴語と呼ぶ）を抽出し，特徴語同士の関連や特徴語グループ同士の関連をユーザにわかりやすい形で表示する。この分析結果から，ユーザは分析対象の大量のテキストに潜む有用な情報を読み取ることが容易になる。

本報告では，情報探索システム“ CoreExplorer ”[1)]を利用したテキストマイニング事例について説明する。

CoreExplorerは，文書の本文中の単語や属性を利用して検索対象を指定したり，様々な観点で絞り込んだ検索を行うことが可能である。また，文書中の単語や属性等を文書を特徴付ける特徴語として選択することもできる。

CoreExplorerの出力情報に，分析クライアントで算出する特徴語間の関連情報等を加えて分析データを生成・表示することでテキストマイニングの機能を実現した。

以下，第２章で本分析システムの動作を説明し，第３章で２つの適用事例について述べる。

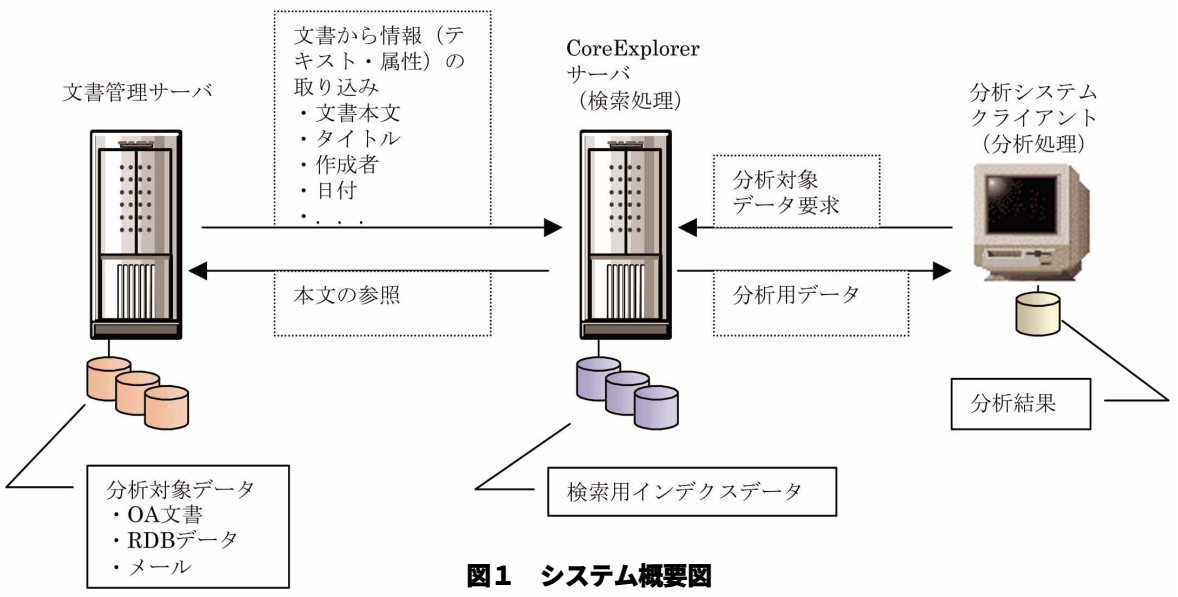

図１　システム概要図

2　分析システムの動作

本システムは，文書の検索を行うCoreExplorerサーバとテキストマイニングを行う分析クライアント（以下，「サーバ」「クライアント」と記述）で構成される（図１）。分析時の動作は，次の３つに分かれる。

⑴分析対象データの要求（クライアントからサーバへの検索要求）
⑵クライアントでの分析処理
⑶分析結果の表示とユーザの操作受付

次節以降，クライアントの分析時の動作⑴～⑶のそれぞれについて説明する。

2.1　分析対象データの要求（クライアントからサーバへの検索要求）

クライアントからサーバに対して分析対象のデータを要求・取得する処理である。

以下の３種類を分析対象の指定に用いている。

①「属性」：文書に付属する属性データ
製品名や顧客名，アンケート記述者の性別や住所（地域）などである。分析対象がリレーショナルデータベース（RDB）の場合は，分析に使用するテーブルのカラムが，OA文書の場合は作成者や最終更新日等が属性となる。文書の絞込みに用いる。

②「単語」：文書中の単語
抽出したい情報に関する１つ，または複数の単語である。その単語が存在する文書の絞込みに用いる。

③「出力特徴語」：特徴語を抽出する対象（属性または単語）
特徴語を文書のどこから抽出するかを指定する。例えば，文書中から特徴語を抽出すれば文書を特徴付ける単語が，属性から特徴語を抽出すれば顧客名や製品名が特徴語となる。

検索サーバとして“CoreExplorer”[1)]を利用した結果，検索条件の設定と同様に分析条件を自由に設定することが可能となった。属性や単語での絞込み条件をゆるめることで文書群全体を広く分析したり，条件を狭めることにより特定の条件を満たす文書群について分析することができる。また，出力特徴語を変更することもできるため，違った側面でデータを抽出することができる。

クライアントからこれらの条件をサーバに送信し，サーバから特徴語一覧，文書一覧，文書間の関連，特徴語を含む文書一覧情報を受け取る。

2.2　クライアントでの分析処理

サーバから受け取る情報（特徴語一覧，文書一覧，文書間の関連，特徴語を含む文書一覧）と，以下４種類のクライアントで生成する情報をもとに分析データを生成する。

・文書のグループ（文書間の関連から）
・特徴語間の関連（特徴語を含む文書一覧から）
・特徴語のグループ（特徴語間の関連から）
・文書に含まれる特徴語一覧（特徴語を含む文書一覧から）

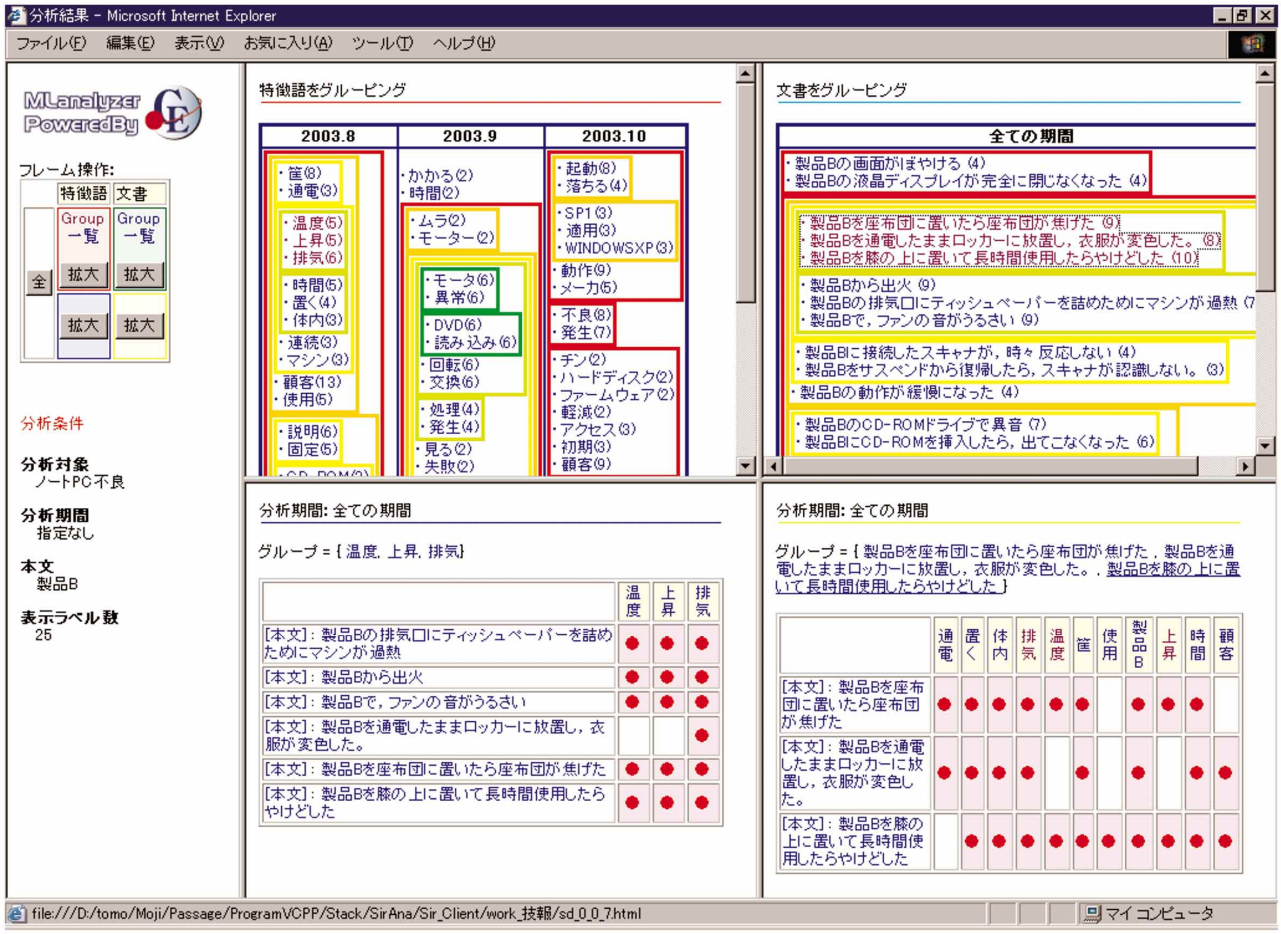

図２　分析結果（保守日報例）

文書のグループは，文書間の関連から求める。文書間の関連が強い文書同士を1つのグループとし，関連の強いグループもそれらのグループを包含するグループとする（図２中上，右上の枠がグループを表す）。

文書間の関連の強さは，文書中の特徴語の重要度をベクトルの要素とする文書ベクトルを用いて，文書ベクトル同士のコサイン角の大きさで表す[1)−4)]。この値はサーバで分析対象データから検索用の索引データを作成する際に算出済みである。

特徴語間の関連も文書間の関連と同様に，特徴語を含む文書（中の特徴語の重要度）をベクトルの要素とした特徴語のベクトル[5)]（特徴語がどの文書に含まれているか）を用い，特徴語ベクトル同士のコサイン角で表す。

特徴語のグループも文書のグループと同様に特徴語間の関連から作成する。特徴語ベクトルのコサイン角が大きいほど特徴語間に強い関連があるとみなしグループ化する。

文書に含まれる特徴語一覧は，特徴語を含む文書一覧の逆写像として求める。

2.3　分析結果の表示とユーザの操作受付

分析結果として特徴語と文書の一覧情報・グループ情報をHTMLのハイパーリンクで表現する（図２中上が特徴語グループ，右上が文書グループ）。一覧情報とグループ情報は，図２左のメニュー画面で表示を切り替える。これらグループ情報や一覧情報中の，特徴語や文書タイトル等をクリックすると，選択した特徴語グループを含む文書一覧（図２中下）や，文書グループに含まれる特徴語一覧（図２右下）へ表示が切り替わる。

特徴語一覧は，より多くの文書に含まれている順番に並んでいる。これにより，例えば月ごとに出力された特徴語一覧から，特徴語の月ごとの推移がわかる（3.1節）。

特徴語のグループ表示では，ある特徴語がどの特徴語と関連があるのかを知ることができる。この結果から，どのような話題が存在するのかを知ることができる。また，文書のグループ表示では文書間の関連を知ることができる。この結果から，同様な内容の文書のまとまりを

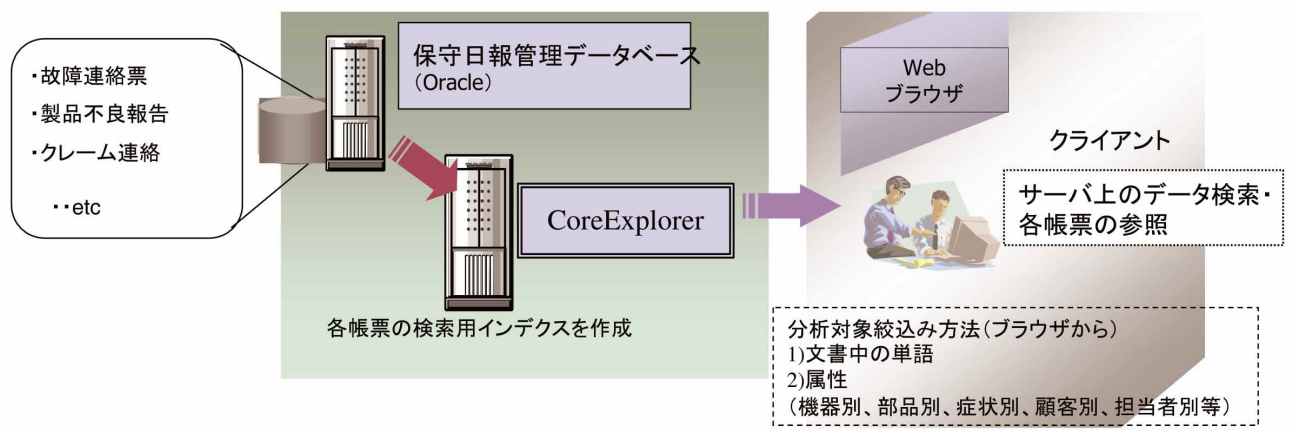

図３　保守日報分析

知ることができる（3.2節）。

3　適用事例

本分析システムを適用した２つの事例について説明する。保守・メンテナンスデータの分析と，複数のメーリングリスト同士の関連を見つけるメーリングリスト分析である。

3.1　保守日報分析

機械のメンテナンス情報である保守日報から，期間の移り変わりによる不良・故障トレンドを分析したい要望があった（図３）。

3.1.1　入力データ

日々保守員から送られてくる保守日報は，２種類（故障事例の日報６万件と製品自体の不良日報２千件）である。保守日報はRDB（Oracle）で顧客名や機種コード，不良内容などの項目でテーブルに蓄えられている。テーブル中の分析に使用するカラムをタグ名としたXMLファイルを分析対象とした。

3.1.2　分析処理

保守日報を期間ごとに分析した。指定期間内の一月ごとを絞り込み範囲とし，月単位の特徴語の変動を出力した。他の絞込みには文書中の単語と機種コードや部品コードなどの属性を用いた。

特徴語が月ごとに並んだ表で表される月別の特徴語の表示（図２中上）から，ある期間に特徴的に出現する特徴語がわかる。さらに，分析条件として機種名等で絞り込むことで，ある期間にある機種に特徴的に出現する特徴語のトレンドがわかる。

3.1.3　分析結果の利用

月別の特徴語の一覧から，その期間に特徴的に現れる（回数が多い，特定の文書に多く現れるなど）単語がわかる。この情報から，シーズンごとの故障の変動や，新製品納入後の故障の傾向を知ることができる。

特徴語とする語を文書中の単語ではなく属性（例えば部品のコードを出力特徴語として選択）とすると，共通部品で起こる不良を調べることができる。この情報から同様の不良を起こす部品グループもわかる。

属性のうち「その他」に分類されるものがある場合，「その他」の文書から特徴語を取ることで，新たな分類作成のヒントを得ることが可能である。将来的には，新しい分類を自動作成することも考えられる。

3.2　メーリングリスト分析

メーリングリストの数が100近く存在する研究所で，「同様の話題を行っているメーリングリスト同士が存在するのかメーリングリスト同士の関連を見たい」という要望があった（図４）。

3.2.1　入力データ

100個のメーリングリスト中，公開されている約40個の各メーリングリストを入力元データとした。分析の前処理として，それぞれのメーリングリストから次の３種類の分析対象データを作成した。

①「メール単位」：メール１件を１文書としたデータ。
　メール中から【本文】，【送信者】，【送信日】，【タイトル（サブジェクト）】を抽出した。

②「月単位」：メール一月分を１文書としたデータ。
　一月分のメールの本文を連結して【本文】とし，

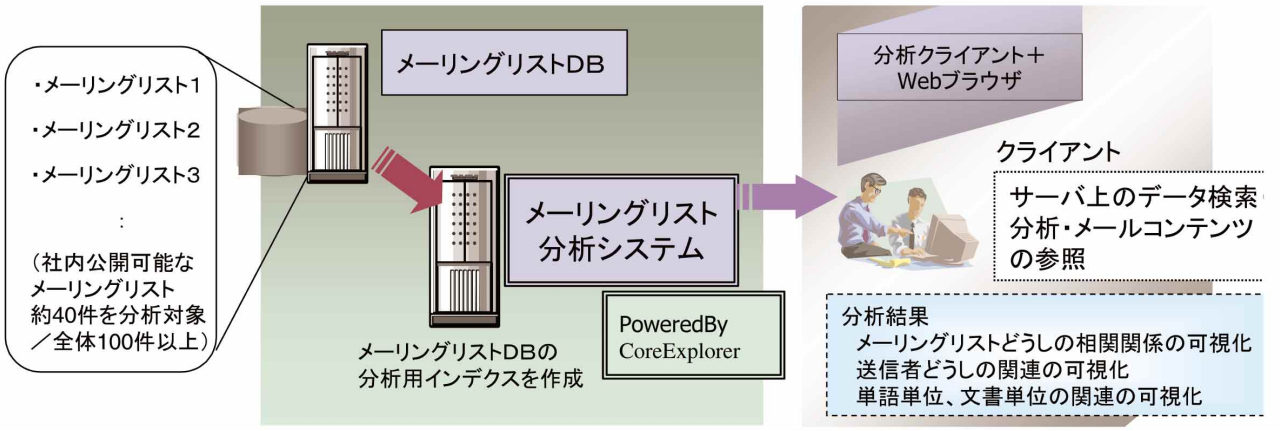

図４　メーリングリスト分析

【送信者】としてメールを送信した人全員を，【送信日】を最新の送信日に，【タイトル】をメーリングリスト名＋年月にしたデータを作成した。

③「全メール単位」：メーリングリスト内全てのメール本文を結合して１文書としたデータ。

メール量が多く本文が長くなる場合は，最新のメールから一定量を制限として【本文】を抽出した。

3.2.2　分析処理

分析対象の絞込みには，メール本文中の単語（【本文】の内容）や，【送信者】や【送信日】などの属性が利用できる。出力特徴語としては，メール本文中の特徴的な単語（【本文】の内容）や，送信者などの属性を指定できる。

メーリングリスト中の全メールを１つの文書と見立て全メール単位で分析したところ，文書のグループ表示（図２右上に相当）から共通の話題を持つメーリングリスト同士がグループ化されることが確認された。グループを選択すると，メーリングリストのグループとそれらのグループに含まれる特徴語の対応が表となって表示され（図２右下に相当），どのような内容で関連しているのかが分かる。出力する特徴語として送信者を選択すると，関連するメーリングリストに関わっている送信者一覧を知ることができる。また，関わるメーリングリストが似ている送信者のグループを知ることができる。

3.2.3　分析結果の利用

「全メール単位」の分析により，メーリングリスト同士の関連を視覚化することができた。この結果から，メーリングリストの統廃合や，メーリングリスト参加者への関連メーリングリストの紹介などを行うことができる。

特徴語として【送信者】を選択することにより，どのような人たちがどのような話題（知識）を持つかがわかる。これを社内の知識マップとして活用することも期待できる。

全メール単位より話題の単位を細かくするのに，月単位での分析も有効である。メーリングリストで様々な話題がある場合，全メール単位では個々の話題が他の話題に隠れてしまう。月単位にすることにより多くの話題を含む状態を回避することができ，共通の話題をもつメーリングリスト同士の関連が見やすくなる。

メーリングリスト間の関連やメーリングリストの内容を表す特徴語について，メーリングリスト管理者の認識と適合しているか現在評価中である。

4　おわりに

文書間・特徴語間・文書と特徴語間の関連を，文書の属性や文書中の言葉で分析を行うテキストマイニングシステムと２つの事例について説明した。

冒頭に述べた通り，大量に蓄積された文書からの知識の抽出・アンケートやクレームからの顧客意見の抽出など，テキストマイニングの適用場面は多い。

本システムでも様々な要求に応えられるよう今後も研究を重ねていく所存である。現在のところ，以下の改良を行う予定である。

4.1　顧客要求に応じた複数の分析手法の提供

アンケートへの適用を考えた場合，「何がどうだったのか」といった情報を抽出したい要求が存在する。本システムでは，特徴語として名詞や形容詞などの単語を抽

出するが，特徴語同士の係り受けの関係は人間が推測することになる。そこで，分析に係り受け情報も利用し，製品の評判や，評判のよい製品などの情報を得る改良が考えられる。

4.2　分析結果の可視化ツールとの連携

本システムでは，構造化されていないテキストデータから，意味のあるデータを抽出して可視化する一連の処理を行っている。データの抽出方法には，始めから分析対象を絞り込んでおく方法と，分析をしながら様々な観点でデータの絞込みを行う方法の２通りがあるが，本システムは始めに分析条件を指定する前者の方式を取っている。

一方，数値データを扱うデータマイニングシステムでは，あらかじめ分析用のデータを一括して保持しておき，クライアント画面で表示するデータを動的に変更しながら分析するシステムが存在する[6)]。この方式の利点として，その都度サーバに分析結果取得の要求をしないため高速に分析できる点がある。また，クライアントでインタラクティブに視点を変更しながら分析を行うGUIメニューも豊富に用意されている。

本分析システムにおいても，テキストデータからデータマイニングに利用できる定量的なデータを一括して抽出・生成しておき，既存のデータマイニングシステムを利用してインタラクティブに視点を変更しながら分析する連携方法が考えられる。

本論文で述べた分析システムは，顧客先で適用評価中である。今後も改良を重ね，様々な顧客要求を満足するシステムに発展させる所存である。

参考文献

1）塚原朋哉　他：情報の体系化と視覚的検索方法，日立TO技報第８号，pp.5-12，2002
2）高梨勝敏　他：ミクロ・コミュニティの知識交流システム「inxs」，日立TO技報第７号，pp.64-70，2001
3）鷹尾誠一：ニュース音声に対する検索方法の比較，信学技報，NLC99-41，pp.97-102，1999
4）木本泰博：意味属性を基底とするベクトル空間法の検索精度，信学技報，NLC99-24，pp.37-44，1999
5）高野明彦：汎用連想検索エンジンの開発と大規模文書分析への応用，情報処理振興事業協会，2002
6）BIソリューション　PowerPlay： http://www. hitachi-to.co.jp/rod/prod_2/power/top.html

塚原　朋哉　　1997年入社
ナレッジソリューションＳＢＧ
CoreExplorer・テキストマイニングツールの研究開発
tomo@hitachi-to.co.jp

佐藤　俊也　　1993年入社
ナレッジソリューションＳＢＧ
知識交流システム・CoreExplorer・テキストマイニングツールの拡販，コンサルティング
shu_sato@hitachi-to.co.jp

椿山　俊和　　2001年入社
ナレッジソリューションＳＢＧ
CoreExplorerの開発・拡販
tsubaki@hitachi-to.co.jp

高梨　勝敏　　1995年入社
ナレッジソリューションＳＢＧ
知識交流システム・CoreExplorerの開発・販売
takana@hitachi-to.co.jp

井上　　悠　　2002年入社
ナレッジソリューションＳＢＧ
知識交流システム・CoreExplorerの開発・販売
haruka-i@hitachi-to.co.jp