CoreExplorer を活用したテキストマイニングソリューションの開発

Development of Text Mining Solution Using CoreExplorer

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

テキストマイニングとは，自由記述文のテキスト情報から，内容の傾向や言葉の相関関係を抽出する手法である。本報告では，テキストマイニング機能を有する製品である CoreExplorer を活用したテキストマイニングソリューションについて紹介する。従来，同製品の機能は出力される特徴語の頻度情報からの分析のみであったが，文書の著者と文書が属する技術など異なる属性の特徴語どうしの関連を分析する機能や，特徴語をカテゴリにまとめる機能を新たに開発した。分析結果の出力には，数値化された分析結果を表形式に出力する機能と，階層的に定義された複数のカテゴリに属する特徴語を視覚的に表現するビューワを用意した。これらの機能を選択・組み合わせることにより，異なる分析要求に適用した事例を示す。

１．はじめに

アンケートの自由記述回答部分や月報のコメントなど，自然文であるテキスト情報は回答者が自由に記述することから，選択式の項目や数値データよりも具体性の高い意見を抽出できる可能性がある。テキストは記述者の真意を知るための貴重な情報源である。

(株)日立東日本ソリューションズでは，情報探索システム”CoreExplorer” [1) 2)] を利用したテキストマイニングツールを開発し[3)]，他社文書管理部門などへのテキストマイニング環境構築やテキストマイニングコンサルティング業務を行っている。

テキストマイニングツールは答え(問題に対する回答)を出すわけではない。テキストマイニングの役割は，分析者が分析対象から意味を汲み取りやすいように特徴的な単語や意味のある文を抽出したり，それらの文書中での頻度や重要度などの統計情報を指標として抽出することにある。それらの統計情報を用いて人間が解釈を行う。どのような分析を行うかによって分析に必要な情報が異なり，情報を抽出するテキストマイニングの技術も異なる。

以下，２章ではテキスト分析に必要な情報について説明し，３章ではそれらの情報を活用した CoreExplorer のテキストマイニングソリューションの開発について述べる。最後に開発したソリューションの適用事例を紹介する。

２．分析に必要な情報

2.1　分析の単位となる要素

テキストマイニングでは，文書からある単位で分析に用いる要素を取り出す。要素の候補としては以下の６点があげられる。

(1)　文書の内容を特徴付ける特徴的な単語
(2)　修飾関係にある 2，3 単語からなる係り受け情報
(3)　意味的なまとまりである短文や主語・述語・目的語などの単語の組

(1)-(3)の情報は文章中から抜き出される情報であり，後者になるほど分析結果が意味的にわかりやすい情報になる反面，高度な日本語処理や定義辞書などの準備が必要になる。

(4)　顧客名やクレーム ID などの文書の属性

文書につけられた属性(RDB のカラムで管理されるような情報）である。アンケートの分析において年代による違いや性別による違いを調べるときや，クレームの分析において機種ごとの比較をするような場合に使用する。

(5) 文書 ID など文書そのものをあらわす属性

文書 ID など，文書の内容である文章ではなく，文書を表す概念的なものである。類似する文書どうしの識別子として使用する。

(6) (1)-(5)の各要素をまとめたカテゴリや，カテゴリをまとめた階層的なカテゴリ

要素をある規則でまとめたカテゴリである。単語をカテゴリにまとめる場合，カテゴリは上位概念等を表すシソーラスなどにあたり，抽出された単語が属するカテゴリの偏りから分析を行う。同様の意味を持つ複数の単語を 1 つのカテゴリとして定義することで，単語よりも意味（カテゴリ）の頻度による分析が可能となる。

2.2 要素の数値化

2.1 節の要素は定性的な情報であり，どんな要素が存在するのか，といった分析が可能である。例えば，“あるアンケートでは「日立 TO」という言葉が出現している”などという情報がわかる。

これら定性的な情報に文書中での重要度や出現文書数などの統計量を付けて数値化することで，どんな情報がどれくらいあるのか，といった定量的な分析が可能になる。定性的な情報に付与する数値を以下に分類する。

2.2.1 単一要素の数値化

ある要素が分析対象の中で，どの程度あるのか，どの程度重要なのかを調べるために必要な数値である。付与する数値には以下がある。

(1) 出現文書数（要素が出現する文書数）

この頻度によって，アンケート中にどのような意見（要素）が何件の回答にあるか，などを調べることができる。

(2) 文書中での重要度

分析対象中の，ある属性を持つ集団においてのみ特徴的な意見の抽出や，回答数は少ないが注目すべき意見の抽出などに重要度の考え方を利用できる。

例えば，1000 件中 5 件の文書にしかその要素が現れないが，要素の重要度が高いと判定された場合に少数意見として抽出するといった使い方がある。

2.2.2 複数要素の関連度の数値化

複数要素どうしの関連度を付与する数値には以下がある。

(1) 特徴語どうしの関連度

特徴語どうしの関連の強さ（関連度）は，同じ文書にともに出現する特徴語どうしで高くなる値である。例えばクレームとの関連度から，ある種のクレームの多い製品を調べることなどに利用できる。文書に「人」と「技術」という属性があるとき，これらの属性から特徴語を抽出すれば，技術を持つ人の分析等に利用できる。

(2) 文書どうしの関連度

文書どうしの関連度は，文書の内容が似ているほど（文書に出現する単語が同じほど）高くなる値で，分析対象の文書群を大まかに分類するための文書のグループ化に利用できる。

2.2.3 時系列での数値化

単一要素・複数要素の数値を，分析対象期間別に数値化した情報である。月毎・日毎などで，どんなクレームが増加傾向にあるのか，また，突然件数が増えた故障日報はないか，といった分析に利用する数値である。

2.3 数値化された情報の出力

数値化された要素情報を外部出力，および可視化する機能である。単一要素と時系列の値を，項目を列挙する軸か時間軸のグラフとして表現したり，複数要素の値を縦軸と横軸に項目を列挙したマトリクスとして表現したりすることができる。また，項目間の関連度を距離として 2 次元平面に項目を配置した CoreExplorer の特徴語マップのような形で表現することもできる[1) 2)]。

３．テキストマイニングソリューションの開発

3.1 分析の単位となる要素の抽出

CoreExplorer によるテキストマイニングの分析で扱う要素は，2.1 節の(1),(4),(5),(6)の情報である。以下に再掲する。

(1) 文書の内容を特徴付ける特徴的な単語

文章中から抽出する情報のうち，CoreExplorer ではより生に近いデータが扱える単語情報（文書の内容を特徴付ける特徴的な単語）を用いている。単語の切り出しには文章中から名詞や動詞を抜き出すツール[4)]を利用している。分析の際に抽出する単語は，切り出された単語のうち，分析対象として絞り込まれた文書群を特徴付ける文書中の単語である[1) 2)]。

(4)顧客名やクレーム ID などの文書の属性

(5)文書 ID など文書そのものをあらわす属性

(6)(1)-(5)の各要素をまとめたカテゴリや，カテゴリをまとめた階層的なカテゴリ

(6)のカテゴリ分析では，あらかじめ利用するカテゴリの情報を定義しておく必要がある。カテゴリの準備には文書をサンプリングし，トピックを表す特徴語を既存カテゴリに分類したり新規カテゴリを作成し分類する。一般的なカテゴリの定義には既存の定義情報[5)]などが利用できるが，専門的な内容を含む文書の分析になるほど分析対象に特有の知識を用いて作成したカテゴリ定義[6)]が必要になる。

カテゴリの定義は大分類・中分類など階層的に定義することが可能であり，複数の親を定義する（複数のカテゴリに属すると定義する）こともできる。特徴語として文書中の単語のうち形容詞をカテゴリに分けると，「よい」「悪い」といった形容詞の大分類を親として，「見た目」や「操作性」などの中分類に，「きれい」「美しい」などの特徴語を割り当てることで，ある製品で検索した結果，肯定的な意見がどのくらいあるのか，などと分析することが可能になる。

3.2 要素の数値化

3.2.1 単一要素の数値化

単一要素として付与される値は，分析対象をある条件で絞り込んだときに，抽出される特徴語に付与される数値（出現文書数・文書中での重要度）である。

出現文書数は，絞り込まれた文書群のうち特徴語を含む文書数である。

文書中での重要度は，検索キーに該当する単語の文書中での重要度を元に算出される検索結果の文書スコアと，文書中の単語の重要度（tf-idf 法で算出）との積を，検索文書で加算した和（単語スコア）である。

(単語スコア) = Σ(文書スコア)*(単語の重要度)

文書中での重要度としては，上記の数値の他に，複数の分析対象があるとき，他の分析対象と比較した単語スコア（差分単語スコア）を考えることもできる。

(差分単語スコア)

=(単語スコア)－Σ(他分析対象の単語スコア) / N

ここで，N は比較相手の分析対象の数。

差分単語スコアを特徴語の重要度として用い，重要度が負の特徴語を分析対象の特徴語から除くことで（または集合の差分として除くことで），差分単語スコアが正の特徴語（差分特徴語と呼ぶ）はより分析対象を特徴付ける単語と考えることができる。また，共通に現れた単語が双方を共通に特徴付ける特徴語（共通特徴語と呼ぶ）となる。例えば，「犬」や「猫」で検索を行ったとき，「ペット」という単語は共通単語に，「番犬」は猫に対する犬の差分特徴語になる（図 1）。

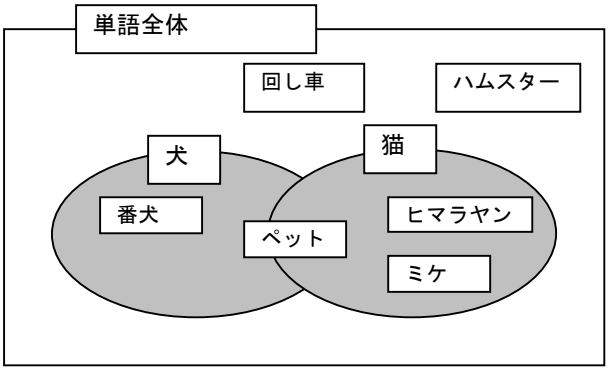

図１ 差分・比較特徴語

差分特徴語により以下のような傾向を知ることができる。

- 一般的な頻出単語に隠されてしまっていた少数意見を分析の対象にすることができる
- （時系列で比較することにより）他の時期に比べてある時期で突出している単語を把握できる

CoreExplorer では以上の，出現文書数，重要度，差分特徴語の重要度，の 3 つの数値を切り替えて特徴語に付与することが可能であり，顧客の要求に応じて適した数値を用いている。

3.2.2 複数要素の数値化

特徴語どうしの関連度は，各特徴語を，それらが属する文書の一覧をベクトルの要素としたベクトル形式であらわすことにより，ベクトル空間での距離（コサイン角）を関連度としている。文書どうしの関連度は，文書に含まれる特徴語を要素とするベクトルを考え，特徴語どうしの関連度と同様に求める。

関連の強い特徴語どうしから，例えばクレームの多い製品などの分析が可能になる。

3.2.3 時系列での数値化

単一要素・複数要素の数値を，分析対象期間を絞り込み，複数回分析することで時系列の数値情報が取得可能である。

3.3 数値化された情報の出力

3.2 節の数値化された情報の出力であり，多くの場合グラフや表形式で出力可能である。3.2.2 節の複数要素の関連度の表現で階層的なカテゴリが定義されている場合，階層関係を意識したまま各カテゴリと特徴語どうしの関連度合いを把握することが難しいため，親カテゴリと子カテゴリを展開しながらカテゴリと特徴語との関連を視覚的に把握できるビューワを開発した（図 2）。

図２ 特徴語のカテゴリ表示

配置する特徴語の座標（ベクトル l_i）を以下の式で求める。

$$l_i = \sum_j R_{ij} * C_j$$

ここで，C_jは表示されている j 番目のカテゴリの座標であり，初期位置はランダムな順番で画面の中心の円の円周上の等間隔の位置である。R_{ij}は i 番目の特徴語と j 番目のカテゴリとの関連度である。関連度は 3.3 節の特徴語どうしの関連度と同様に，特徴語・カテゴリが属する文書を要素とするベクトルのコサイン角で求める。

特徴語とカテゴリとの関連度合いを正確に表現するためには表示する表示するカテゴリの数だけ軸が必要であるが，多くのカテゴリと関連のある特徴語は中心の座標付近に配置され，特定のカテゴリと関連の強い特徴語はそのカテゴリ付近に配置されるため，二次元空間での表現や，初期位置を円周上にランダムに配置しても十分である。

近傍に配置された特徴語に関連のあるカテゴリノードをクリックしていくことで，カテゴリに関連のある情報を調べていくことが可能である。

４．テキストマイニングの実践事例

今回開発したテキストマイニングソリューションを，顧客に適用した事例を紹介する。

4.1　分析の単位となる要素

CoreExplorer では特徴語として文書中の単語や文書の属性など自由に選択可能である。適用事例ごとに適した特徴語を選択して分析を行っている。

あるコールセンタのオペレータ用検索画面では，文書をカテゴリに分類して表示している。検索結果の文書をディレクトリ（階層）に分類して CoreExplorer の GUI に表示させ，利用者の利便性に配慮した（図 3）。

図３ 文書の階層表示

階層情報はコールセンタの文書管理者があらかじめ作成しておいた情報を利用した。オペレータが検索した文書がディレクトリに分かれて検索結果として出力されるため，目的情報を持つ文書の発見の手助けになる情報の提示が可能となった。

4.2　要素の数値化事例

定性的な分析の要素に，定量的な情報を付与した事例について，情報の出力まで含めて紹介する。

4.2.1　単一要素の数値化

複数の属性から抽出した特徴語の一覧を出力する機能をレポート出力機能として実装した。図 4 がレポート出力機能の例である。図中の 4 つの表それぞれが異なる属性を表しており，各属性から抽出された特徴語が表の行に出力される。各行では全文書において特徴語が出現する文書数と，今回の検索において特徴語が出現する文書数が表示される。

図 4 レポート出力機能

ある二輪車メーカでは，営業報告書を分析対象として本機能を適用した。ある車種で分析対象を絞り込み，その車種についての月報を書いている地域ごとの割合や営業所ごとの割合の調査を行った。一覧表の形式で出力されるため，その車種について特出して記述している営業所などの把握が容易となった。

次に，3.1 節の差分特徴語と共通特徴語を表形式で表現した出力例を図 5 に示す。左の 1 列に共通特徴語を，右側の列に他の検索キーに対するそれぞれの検索キーの差分特徴語を重要度順に表示している。

本機能も，先の二輪車メーカにおいて“自社と他社ライバルメーカとの相違点と共通点”の調査に使用した。他社にあって自社にない特徴語が重要度順に表示されるため，営業戦略の違いなどの気付きを助けられた。

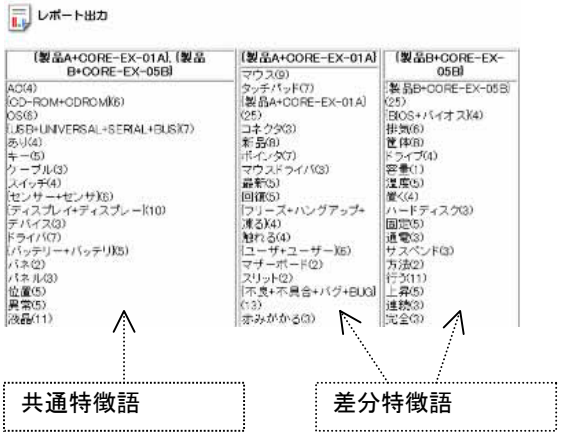

図５ 差分・比較特徴語の表出力

4.2.2 複数要素の数値化事例

特徴語どうしの関連度の適用事例について以下 2 例紹介する。

(1) 特徴語どうしの相関関係出力

(2) Know-Who 分析

(1)の特徴語どうしの相関関係出力では，特徴語どうしの関連度を表形式で表現したものである（図 6）。行・列に抽出された特徴語を配置し，特徴語間の関連度を色付き（赤に近くなるほど関連度が高く，青・白に近いほど関連度が低い）で示す。

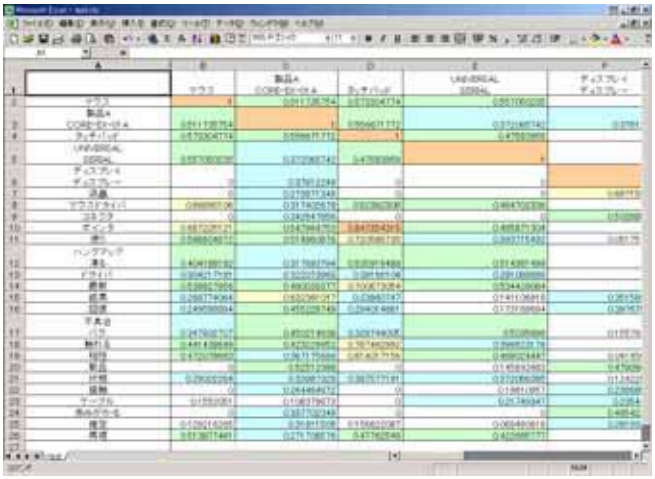

図６ 関連度のマトリクス表

図 6 をグラフ化したものを図 7 に示す。平面の左右の軸に特徴語を配置し，関連度をグラフの高さで表現している。関連度の強い単語どうしの組み合わせ情報を 2.1 節(6)の特徴語のカテゴリの作成の判断材料として適用した。

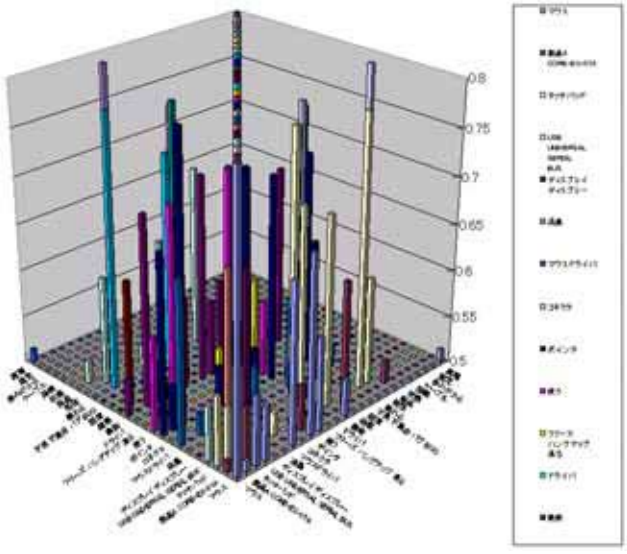

図７ 関連度の 3D グラフ

⑵の Know-Who 分析は，異なる属性の特徴語どうしの関連を抽出する機能である。研究所での論文を対象にし，Know-Who 検索システムとして実装した。

図８　Know - Who 検索

本事例では，論文のデータベースを分析対象とし，論文の著者と論文の技術の間で Know-Who 分析を行っている。ある技術で検索したときに関連する人（図 8 下側の 3 つの表の 3 人）と，それらの人が持つ技術（図 8 表中の「関連キーワード」）を出力する。

また，分析対象のフィールドを「人」と「技術」以外にも設定できるため，興味のあるフィールドどうしの関連を分析することが可能である。「人」「技術」以外の分析では，例えば「クレーム」と「機種名」との関連を調べれば，ある種類のクレームが発生しやすい機種や，共通に発生するクレームどうしの情報を得ることができる。

その他，要素どうしの関連度は，CoreExplorer の基本機能である関連する特徴語どうしを二次元マップ上の近傍に配置する特徴語マップ画面へ適用したり，関連する文書どうしをグループ化することで，類似話題を持つメーリングリストどうしを発見する事例[3)]に適用している。

4.2.3　時系列での数値化

故障日報や月報を分析対象として注目特徴語の出現文書数をデータベースに格納したり，グラフ化する適用を行った。注目単語の設定や分析対象期間の設定などを評価中である。

4.3　数値化された情報の出力

中毒情報の文書群から作成した階層的なカテゴリと検索された文書中の特徴語との関連を 3.3 節のビューワで表示を行った。図 2 がこの表示例であり，中心の赤いノードが親カテゴリ（自然毒）を，周辺の赤いノードが子カテゴリ（植物や咬刺傷）を表しており，各カテゴリに関連のある特徴語がカテゴリノードの近傍に配置され（「植物」カテゴリに「キノコ」,「咬刺傷」カテゴリに「アナフィラキシー」など），どのカテゴリとも関連のない特徴語が画面左上にまとめられている。カテゴリノードをクリックすることでさらに親・子を手繰って移動ができる。また，カテゴリノードをドラッグして位置を変更すると特徴語の位置が再計算される。この表示により，抽出された特徴語がどのようなカテゴリと関連が深いのかを視覚的に理解可能となった。

５．おわりに

本報告書では，これまで開発したテキストマイニングソリューションとその実践事例を紹介した。事例の中で，CoreExplorer の特徴である文書の特徴語の情報や特徴語のカテゴリを定義した情報を用いて，特徴語どうしの関連情報や，頻度情報からさまざまな分析に必要な情報を出力した結果が有効であったことを示した。

しかしながら，現時点ではまだ分析の要素として文書中の単語単位の情報しか扱っていない。そのため，文脈といったより高度な意味解析までの分析が行えていない。

今後は，特徴語として文書中の単語の代わりに短文などより大きな意味の情報の扱いも検討していく。また，現在数値データの分析とリンクした分析機能が無いため，DWH（DataWareHouse）と連動したテキスト分析ができない。例えば，「今月急減した販売台数の原因を探る」といったケースでは，DWH で出力したグラフから直接営業報告書のテキストマイニング結果が出力されるのが理想的であろう。今後はこのような数値情報と組み合わせ分析の自動化も進めていく所存である。

参考文献

１）塚原朋哉，　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）塚原朋哉，情報の視覚的検索方法，fit2003. 一般講演論文集 第二分冊 pp.179-180, 2003.
３）塚原朋哉，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，　日立 TO 技報第 9 号，　pp.54-59，　2003.
４）奈良先端科学技術大学院大学情報科学研究科自然言

語処理学講座『形態素解析システム（茶筌） version 2.3.3 使用説明書』, 2003.
５）日本語語彙大系,NTT, http://www.ntt.co.jp/news/news99/9909/990924.html
６）高梨勝敏，セマンティクス自動抽出によるエンタープライズオントロジの構築，日立 TO 技報第 10 号，pp.33-38，　2004.

塚原　朋哉　　1997 年入社
ナレッジソリューションＧ
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Ｇ
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

渡辺　まり子　　2004 年入社
ナレッジソリューションＧ
CoreExplorer を利用したテキスト分析・コンサルティング
w_mariko@hitachi-to.co.jp

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

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