自然言語処理技術による情報マネジメントの実際

企業における非定形文書の活用促進事例
－営業日報へのテキスト分析技術の適用－

（株）日立製作所 情報・通信グループ営業企画本部
柴田 親男
cshibata@itg.hitachi.co.jp

（株）日立製作所 中央研究所
松田 純一
j-matsud@crl.hitachi.co.jp

（株）日立製作所 中央研究所
小泉 敦子
koizumi@crl.hitachi.co.jp

（株）日立製作所 中央研究所
森本 康嗣
y-morimo@crl.hitachi.co.jp

背景

営業日報は営業活動に伴い発生する企業内文書であるため，ほとんどの企業で作成されている汎用性の高い文書である．紙文書であったり，メール形式であったり，音声情報であったり，週報や月報になっていることもあり，形式はさまざまであるが，最前線の企業活動の最新情報が豊富にそろっているため，うまく活用すれば，有効であることは間違いない．本稿では，テキスト分析の観点から営業日報の分析活用事例および必要技術を紹介する．

表 -1 に典型的な営業日報のデータ形式を示す．案件ID，活動日，訪問先，面会者などの定型データ部分とテキスト自由記述部分とに大きく分けることができる．従来，OLAP ツールやデータマイニングツールを用いて定型データ部分の分析が行われてきたが，豊富な情報，定型データ部分では十分に書けない情報がテキスト自由記述部分に書かれていることが少なくない．営業日報の有効活用にはテキスト自由記述部分の分析が必須である．

フィールド名	データ形式	補足
案件 ID	定型データ（数値）	
活動日	定型データ（数値）	報告内容のアクション日
報告者 ID	定型データ（数値）	社員番号
報告者氏名	定型データ（文字）	
同行者氏名	定型データ（文字）	
訪問先	定型データ（文字）	
面会者	テキスト自由記述	面会した相手
進捗状況	定型データ（数値）	引き合い，見積もり等の状況
結果	定型データ（数値）	受注，失注等結果
報告内容	テキスト自由記述	

表 -1　営業日報のデータ形式

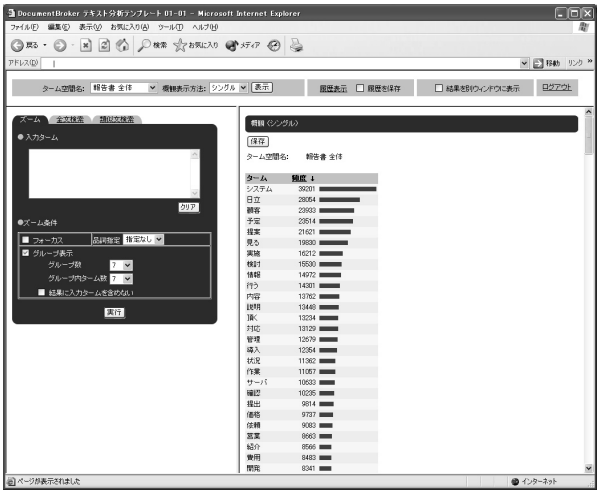

図 -1　単語頻度情報

図 -2 「サーバ」の製品別頻度

テキストの活用目的

テキスト情報は，以下のような目的に使うことができる．

(1) 現状分析，定量化，確認

日々発生する大量のテキスト文書からマネージャが情報収集して，現状分析，確認，定量化を行い，企業活動の戦略策定に役立てる．具体的には，商品企画／改善，営業戦略立案，などである．

(2) 通常と異なる情報がないかチェック

マネージャが，通常と異なる現象（たとえば，事故，不正，急激な売り上げの落ち込み，顧客クレーム，など）を発見し，事前対処，早期対処を可能とする．

(3) 類似案件情報の検索

担当者が，類似案件の検索を行い，自己の業務へのヒントを得るための参考情報として活用する．これには，類似文書検索技術を用いることにより対応できる．

以下では，これまで比較的議論されてこなかった (1) と (2) を中心に述べる．

テキスト分析の機能と事例

テキスト分析の機能

テキストをさまざまな観点から分析するための基本機能として以下の 2 つがある．

(1) 現状分析，定量化，確認のための機能

一般には，ある観点で報告内容を分類・整理し，どんな事象が多いかを把握することが基本となる．同一事象であるかを判断するには，同一の単語や単語群（単語の組）が出現しているかで近似することができる．単語や単語群の出現頻度を調べればおおよその傾向を得ることができる．

(2) 通常と異なる情報がないかチェックするための機能

事故につながる特定のキーワードの有無や (1) で述べた定量化した数値の時系列推移から急激な変化を検出し，通常と異なる現象の可能性をチェックすることができる．

テキスト分析の事例

実際の営業日報データを活用した分析事例を 3 つ紹介する．営業企画本部で，マネージャがツールを活用しながら，業務知識をもとにさまざまな観点から分析を定期的に行っている．分析した結果は，営業担当者や製品開発担当者にフィードバックできるように Web ポータルにのせており，参照，活用できるようにしている．また，必要に応じて，システム設計者が，カスタマイズや辞書構築を行っている．

(1) サーバ関連の商品企画・改善

情報システム関係の営業日報の報告書を対象とした分析事例を紹介する．まず，報告書全体の概要を得るため，単語の頻度情報を見る．図 -1 に示すように「サーバ」という単語が比較的高頻度で出現していることが分かる．サーバの機種ごとに，言及されている頻度を調べると，図 -2 に示すように，HA8000 が最も件数が多い．さらに，HA8000 に関して，仕様，周辺機器，ソフトウェア，他のサーバとの組合せ，用途，競合他社製品，などの観点から顧客の声を分析する．

図 -3 は，「HA8000」の関連用語を分類表示したものである．一番上の分類グループには，ラックやキャビネット，UPS など周辺機器がまとまっている．

図 -3 「HA8000」の関連用語参照

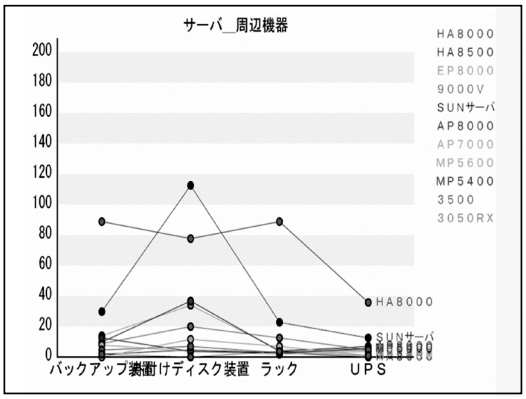

図 -4　サーバ機種別周辺機器への顧客ニーズ

図 -5 「ラック」の関連用語参照

図 -6 「UPS」の関連用語参照

他分類グループも参照した上で，HA8000 の話題としては，バックアップ装置，外付けディスク装置，ラック，UPS の 4 つを主要な話題として取り上げることとする．これらの話題について，定量的な傾向分析を行ってみる．たとえば，図 -4 のグラフは，サーバ機種別にどのような話題が多いかを示すものである．

予想外に多かった周辺機器や他のサーバとの組合せで発生するラッキングの問題に着目する．「ラック」の関連用語の参照をすると，図 -5 に示すように，「UPS」がよく話題に出てくることが確認できる．一方，「UPS」から見ても，図 -6 に示すように，ラックに関連して話題になりやすいことが分かる．

実際の営業日報報告書を検索してみると，「UPS がラックに収まらない」などの原因で予定外の出費や時間の浪費が発生するといった事例が少なくないことや，「スペースの有効活用のため，ラックの数を減らしたいという顧客ニーズ」があることが分かった．このような情報をもとに，以下のような商品開発に結びつけることが可能になった．

・ラックに収めやすい UPS（薄型 UPS など）の開発
・UPS を収めやすいラックの開発

(2) 統合管理パッケージ関連の商品企画・改善

統合管理パッケージ製品として，GEMPLANET という名称の，人事，労務，会計，販売に的を絞った中堅企業向け ERP 製品がある．まず，GEMPLANET が報告書中でどのような用語と関連しているかを調べてみる．関連用語を分類して表示したものを図 -7 に示す．

図 -7 から，以下のようなことが分かる．

・人事，労務，会計，販売のキーワードが出るのは当然であるが，これ以外にも，債務管理，固定資産，連結決算，経営分析というキーワードが出現しており，別の機能としても使われている可能性がある．
・ERP なので，生産管理というキーワードも当然出てきており，UNIMEX という製品と絡めて動向調査が必要である．
・経営分析の観点からは，HITSENSER という製品との関連も調査要である．

次に，定量的な分析を行ってみる．定量的な分析は，

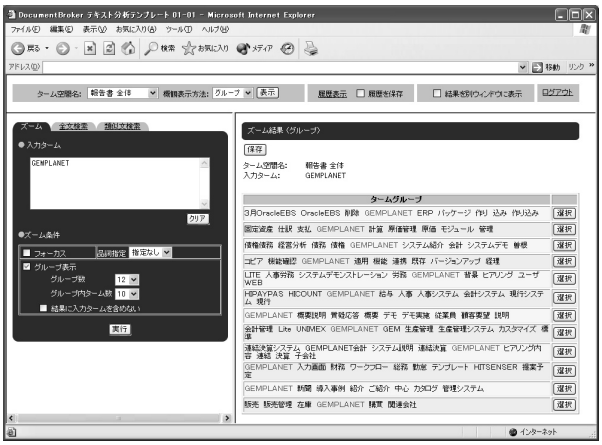

図 -7　GEMPLANET の関連用語

表名: GEMPLANET機能/進捗 | 表示 | グラフ表示 | 更新 | 削除
対象データベース: 報告書
絞込条件: (なし)
縦軸: GEMPLANET_機能
横軸: 進捗_進捗別
"その他"の出力: off

	引合前	引合	見積	受注	提案	内示	アフターフォロー	–
会計管理	94	146	30	10	66	4	1	2
固定資産	16	34	3	2	15	2	0	1
連結決算	25	30	3	0	5	0	0	2
債権債務	11	34	3	2	15	0	1	0
人事管理	67	83	14	7	23	2	3	2
労務管理	49	69	12	5	18	5	3	2
勤務管理	8	25	2	3	5	4	0	0
販売管理	14	12	4	0	5	0	0	1
購買管理	1	1	0	0	1	1	0	0
在庫管理	4	4	0	1	6	1	0	0
生産管理	20	17	5	0	12	1	1	0
原価管理	10	9	0	0	8	0	0	0

図 -8　GEMPLANET の機能－進捗状況別の報告件数

	会計管理	固定資産	連結決算	債権債務	人事管理	…
A 支社	852	82	121	55	415	
B 支社	681	135	104	102	436	
C 支社	398	57	29	57	166	
…						

表 -2　GEMPLANET の支社－機能別報告件数

	活動中	未接触	応答待ち	一時凍結	完了（すべて受注）	完了（一部受注）	完了（失注）
A 支社	188	21	5	0	8	2	2
B 支社	119	17	7	2	1	0	2
C 支社	16	7	0	6	1	0	0
…							

表 -3　GEMPLANET の事業部－進捗状況別報告件数

	2002/1	2002/2	2002/3	2002/4	2002/5	2002/6	…
会計管理	15	20	21	13	29	31	
固定資産	2	5	5	3	3	2	
連結決算	4	1	4	3	12	6	
債権債務	2	3	3	4	7	5	
人事管理	6	15	13	11	11	17	
…							

表 -4　A 支社での GEMPLANET の機能－日時別報告件数

いくつもの観点から，その観点に言及している報告書の数を数値化することによって行う．図 -8 は，縦軸に GEMPLANET の機能，横軸に進捗状況をとった場合の報告件数を表したものである．

会計管理，人事管理，労務管理の機能が販売の中心であることが分かる．販売管理，生産管理の提案もしているが受注は少ない．

このほかにも，縦軸に支社，横軸に GEMPLANET の機能をとった表（表 -2），縦軸に支社，横軸に進捗状況をとった表（表 -3），縦軸に機能，横軸に日時をとった表（表 -4）など，さまざまな表を複合的に参照することにより，支社別活動状況を把握することができる．

このように観点を自由に変えながら定量的な分析を行い，出てきた表やグラフをもとに，必要に応じて報告書本文を参照して，新たな気づきの発見や拡販・社内展開施策の立案に役立てている．

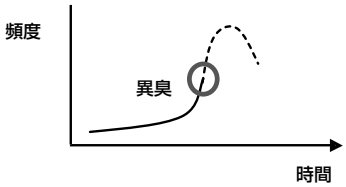

図 -9　単語の時系列頻度

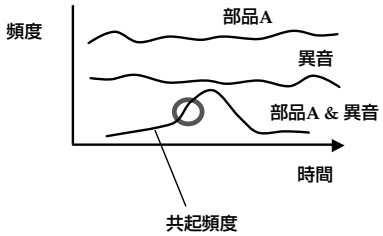

図 -10　単語ペアの時系列頻度

(3) 製品不具合の早期検知

商品 A のサポートに関する営業日報中での単語頻度の時系列推移を見ることにより，製品不具合の早期検知に役立てることができる．製品不具合に関する分析では，不具合に関係する単語を分析の観点とすることができる．不具合に関係する単語の一例として，以下のようなキーワードが挙げられる．

異常，異臭，異音，壊れる，フリーズする

たとえば，図 -9 に示すように，ある時期に「異臭」というキーワードが急激に増えた場合，事故の可能性を疑うことができる．

一方，単一キーワードでは時系列頻度に大きな変化が出ないが，複数の単語の組合せ（共起）頻度をとるとある時期に急激に増える場合もある．図 -10 は，「部品 A」という語と同時に出現する「異音」というキーワードがある時期に急に増えていた例である．「部品 A」「異音」各々のキーワードの出現頻度だけでは特に時系列的に大きな変化が現れないが，組み合わせた場合，特定時期に大きな変化が生じることもある．すべての共起の頻度を調べるわけにはいかないので，部品・構成要素，人，会社名，などあらかじめ分析の観点となり得るキーワードを抽出しておく．

これにより，頻度がピークに達するまでの期間 45 日に対して，15 日ほどで事故発見ができた事例があった．

適用技術

前項で述べたテキスト分析では，以下のような自然言語処理関連技術が必要である．

(1) 日本語解析技術

テキストから単語や単語群を抽出するためには，形態素解析，構文・意味解析の技術が必要である．この技術は，従来から研究開発が進められてきているが，営業日報などの企業内文書には，省略語／表現や誤表記，かな漢字変換誤り，などが多いという特徴がある．短時間で報告書を書かねばならないため，正しい入力を前提とした日本語解析技術では不十分である．たとえば，SANRISE という製品名称に対して実際には，SUNRISE，SANRIZE などの誤表記が少なからず存在する．

(2) 辞書構築

企業内用語が頻出し，また，部署によって呼び方が異なることもあり，汎用形態素解析辞書では未知語が数多く出現する．実際の利用に際しては，企業内用語の辞書構築が必須である．また，省略語に対しては，同義語辞書の充実が必要である．登録候補語自動抽出ツールによる効率的な辞書作成を行っている．

(3) テキストの定量化

単語や単語群の出現頻度を求めるに当たり，いくつかの工夫が必要である．

単語頻度を調べる場合，すべての単語の頻度を求める必要はなく，意味のない単語をストップワードとして除外したり，また，ある観点（分析軸）を定めてその頻度のみを求めることが必要である．たとえば，営業日報では，「見積り」や「説明」などの語は高頻度語であることが当たり前であり，これ自体は特徴となり得ない．分析軸としては，たとえば，商品名，人名，会社名，などのいわゆる固有表現が有効である．固有表現抽出の自動化に関して自然言語処理技術が活用できる．

また，単語群の頻度を調べる場合，単語群のとり方にもいくつかの方法がある．通常，動詞—目的語の組みをとることが多いが，そのほかにも，取り方として

主語—動詞—目的語など別の構文要素の組

形容詞—目的語の組

名詞＋の＋名詞の組

同一文中の単語の共起（構文関係なし）

単語の前後 n 文字／単語以内の単語の共起

などいくつもの方法がある．目的に応じて使い分ける必要がある．

(4) 分析軸の設定

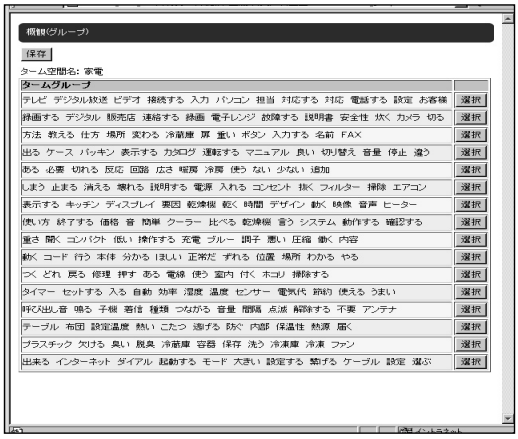

図 -11　テキスト全体の概観表示

分析軸を決めるための支援ツールとして，自然言語処理を活用することができる．分析軸の基本設計は業務的観点から行う．営業日報では，たとえば，「製品分野」「顧客業種」「競合他社」といった分析軸が考えられる．「製品分野」という分析軸に対しては，「ソフトウェア」「ハードウェア」といった分析軸項目が考えられる．「ハードウェア」はさらに，「サーバ」「ハードディスク」などに細分化される．しかし，この分析軸だけでは，漏れや重なりなどが出る可能性があり，これを防ぐために，以下のような機能を活用している．

機能 1：テキスト全体の概観表示

文書全体に出現した特徴的な用語を抽出して表示することによって，文書に書かれている内容の全体傾向を把握する機能である．これにより，分析軸の漏れを防ぐことができる．具体的には，文書全体に出現した用語を関連が強い用語ごとにまとめて表示(クラスタリング表示)する．関連が強いものがまとめて表示されているため，文書全体に出現した話題を大まかに理解するのに向いている．図 -11 に例を示す．

機能 2：関連用語の参照

ある用語と関連が強い用語を表示する機能である．これにより，指定された用語に関してどのような話題の広がりがあるかを推測することができる．たとえば，「製品 A」の関連用語として「製品 B」「C 社」「価格」を発見し，「製品 B が同時に売れているようだ」「C 社が競合他社か？」「価格が話題になっているようだ」などと推測することができる．

まとめと今後の課題

営業日報を例に取り，テキスト分析の活用事例とそのために必要な技術について述べた．自然言語処理技術をベースにしたさまざまな機能を使ったテキスト分析が必要であり，今後とも精度向上が必須であるが，各種機能の充実だけでなく，どんなテキストで，どんな目的の場合には，どの機能をどういう順で利用すれば最も効果が得られるのかという，いわばテキスト分析のノウハウが大事である．ノウハウの蓄積には，多くの事例を扱うことが必要であるが，効率的にノウハウを蓄積するための手段としても，今後，自然言語処理が役立つのではないかと考えている．

参考文献

1）市村，中山，赤羽，三好，関口，藤原：日報分析システムの開発，信学技報，NLC2000-26（2000）．

2）梶，森本，相薗，山崎，飯田，内田：コーパス対応の関連シソーラスナビゲーション，情報処理学会データベースシステム研究会／情報学基礎研究会研究報告，DBS-118-13/FI-54-13（1999）．

（平成 15 年 9 月 9 日受付）