産業応用分野に求められるテキストマイニング技術と適用事例
―実務を意識した定性情報のクラスタリングと可視化技術―
New text mining technology for industrial application
-Clustering and visualization technology for business qualitative information-

株式会社　山武　村上　英治
アドバンスオートメーションカンパニー　Eiji Murakami

株式会社　山武　木幡　真望
アドバンスオートメーションカンパニー　Motomi Kohata

キーワード
テキストマイニング、クラスタリング、リスク管理、ナレッジマネジメント、知識発見

身近に存在する文書から有用な知識や情報を獲得するための技術として新しい文書クラスタリング手法を提案し、文書解析ツールとして開発したので報告する。提案手法は複数文書から単語-文書行列を生成し、これを単語と文書の2組の集合から成る2部グラフとして扱い、グラフ理論でしばしば利用される2部グラフの分解アルゴリズム（Dulmage-Mendelsohn（DM）分解法）を適用し、単語と文書の2部グラフに対して、グラフを縮約しながら繰り返し適用することで、文書集合の階層的なクラスタリングを実現する。提案手法を適用した文書解析ツールは文書クラスタリングにより大量文書の情報を集約および要約を行い、可視化情報と要約文を生成することを特徴とする。

In this paper we put forward a new document clustering technique as a technology for acquiring valuable knowledge and information from documents close at hand, and we report on our successful development of the technique as a document analysis tool. The proposed technique enables the hierarchical clustering of document sets. In the technique, a term-document matrix is generated from multiple documents; this is then treated as a bipartite graph comprised of the aggregate of 2 pairs of terms and documents; a decomposition algorithm (Dulmage-Mendelsohn (DM) decomposition method), which is sometimes used in graph theory, is applied; and the algorithm is repeatedly applied to the term and document bipartite graph while the size of the graph is reduced. The features of a document analysis tool, to which our proposed technique is applied, is that it can consolidate and summarize even greater volumes of document information in document clustering, and that it can generate visualization information and summaries.

1. はじめに

パソコンやITシステムの低価格化に伴う普及によりお客様相談センタに寄せられるVOC（Voice of Customer）や組織内で作成される営業報告書、またものづくりの現場で作成される品質記録や設備保守記録などのテキスト情報が身近に存在するようになった。

このような状況から、近年はテキストマイニングと呼ばれる膨大なテキスト型情報から有用な知見を得えるための技術や方法論が確立されつつある[(1),(2)]。

本論文では、会社などの組織において身近に存在する文書から有用な知識や情報を獲得するための技術として新しい文書クラスタリング手法を提案し、文書解析ツールとして開発したので報告する。

提案手法は複数文書から単語-文書行列を生成し、これを単語と文書の2組の集合から成る2部グラフとして扱い、グラフ理論でしばしば利用される2部グラフの分解アルゴリズム（Dulmage-Mendelsohn（DM）分解法）を適用する。

DM分解は組み合わせ論的アルゴリズムの中では高速であり、これを単語と文書の2部グラフに対してグラフを縮約しながら繰り返し適用することで、文書集合の階層的なクラスタリングを実現することが可能である。

DM分解を使用する利点は単語と文書の集合の要素数が少数の場合においてもクラスタリングを実施できる点である。例えば企業内で作成される技術文書のように、ひとつの文書が短い場合や文書数が少ない場合においても本手法を適用することができる。一方DM分解の単純な適用では単語のクラスタリングが可能であることが知られているのみである[(3)]。本論文では繰り返しDM分解による文書クラスタリングを提案する。

提案手法を適用した文書解析ツールは、文書クラスタリングにより大量の文書情報を集約および要約し、可視化情報と要約文を生成することを特徴とする。

本論文の構成は、次のとおりである。第2章に提案する文書クラスタリング手法について論じる。第3章は提案手法を適用した、文書解析ツールについて述べる。第4章に適用例につい

て述べる。第5章にまとめを述べる。

2. 提案する文書クラスタリング手法

システムの処理過程と構成を図1に示す。最初に形態素解析[4]により文書から単語を取り出し次に単語-文書行列を作成する。この単語-文書行列に対して繰り返しDM分解を実施する。

また、繰り返しDM分解の際にクラスタラベルとして得られる単語情報を使って重要文を決定し、その文を抽出することで要約文を生成する。

以下、その内容を詳述する。

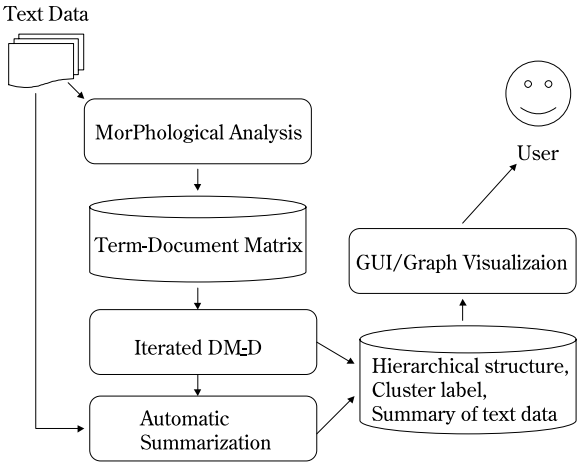

図1. システムの構成

2.1.DM分解によるクラスタリング

文書中に出現する単語間の係り関係などの情報を使って単語を分類することによって、意味素性を抽出する研究が行われている[5]。この研究では単語間の修飾、被修飾の関係を図2のような2部グラフで表し、グラフ・ネットワークの数学的手法であるDM分解として知られる2部グラフの分割手法を利用して単語のクラスタリングを行う[6],[7],[8],[9]。

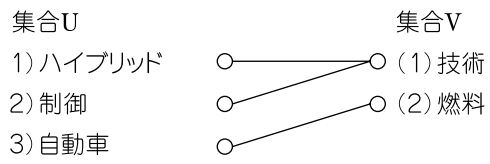

図2. 単語間の結合関係

DM分解は2部グラフGの最大マッチングMを求め（図3(a)）、Mのエッジの逆向きのエッジを付加して(図3(b))、強連結成分$G_{(i)}$（図3(c))を求める手法である。

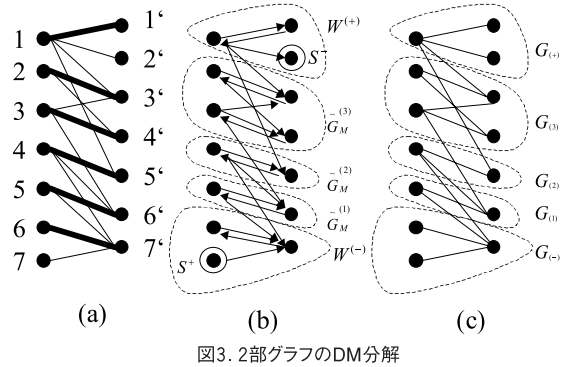

図3. 2部グラフのDM分解

自然言語処理において、単語と文書の関係は単語-文書行列で表現される。これはある単語が文書に出現する場合は1とし、出現しない場合は0とするデータ行列である。この0-1行列から文書および単語をグラフのノードとする図4（左）のような2部グラフを得る。

この2部グラフに対してDM分解を実施することで、文書のクラスタリングが可能となる(図4（中）)[10]。

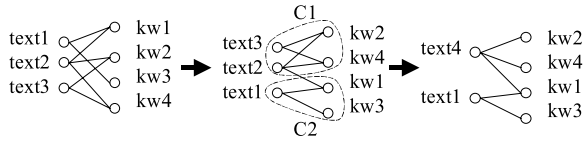

図4. 単語-文書行列のグラフ表現
（左はDM分解前、中はDM分解後、右は統合文書生成）

2.2.繰り返しDM分解による文書クラスタリング

この節では繰り返しDM分解によって階層的なクラスタリングを行う方法について述べる。

最初にDM分解によって文書のクラスタリングを行う。次にクラスタリングされた複数の文書を統合し仮想的な新しい文書を生成する。同時にクラスタリングされた文書を2部グラフから消去する。これを繰り返し実行することで文書クラスタリングを行う。例えば図4（左）の単語-文書の関係がある時、DM分解を実施することで図4（中）を得る。この結果から単語-文書行列においてtext2、text3から新しく統合文書text4を生成する（図4（右））。

文書の特徴量を文書のキーワードとした時、同じクラスタに分類される複数文書をひとつにまとめた統合文書のキーワードは元の文書のキーワードの和集合となる。

このようなサンプルから仮想的な代表サンプルを生成してクラスタリングを行う方法は、CURE(Clustering Using Represent-atives)[11]でも見られるが本手法はサンプルを2部グラフ構造のデータとして取り扱うところが特徴である。

本手法は仮想的な代表サンプルの生成を繰り返し行うことでボトムアップ的な階層化クラスタリングを実現する。

今までの内容をまとめると手順は次のようになる。

(1) 単語-文書行列をDM分解し、2部グラフGの強連結成分Hを得る。
(2) Hにクラスタリングされた複数の文書から統合文書を生成し、2部グラフGを変形しG'とする。
(3) 構造が変わった新たな2部グラフG'に対してDM分解を行い強連結成分H'を得る。
(4) (2)、(3)を新しい強連結成分H'が得られなくなるまで繰り返す。

単語-文書行列に対して最初のDM分解で得られるクラスタリング結果をstep0での結果として記憶する。以降DM分解を繰り返して得られるクラスタリング結果とDM分解の繰り返し数(step数)を記憶する。これをDM分解によって新しいクラスタリング結果が得られなくなるまで実行する。この計算手順の繰り返し数はボトムアップ的な階層化クラスタリングにおける階層情報として使う[12]。

図5のstep0で文書aと文書bが同一クラスタに分類されstep1で統合文書と文書gが同一クラスタに分類されたときにこの二つのクラスタの間に線を引き、統合関係を表す。繰り返し

DM分解の結果、図5のような階層構造が得られる。

ここで、a,b,...,kは与えられた文書を表し、ⅰ,ⅱ,...,ⅵは繰り返しDM分解によって得られた統合文書を表す。

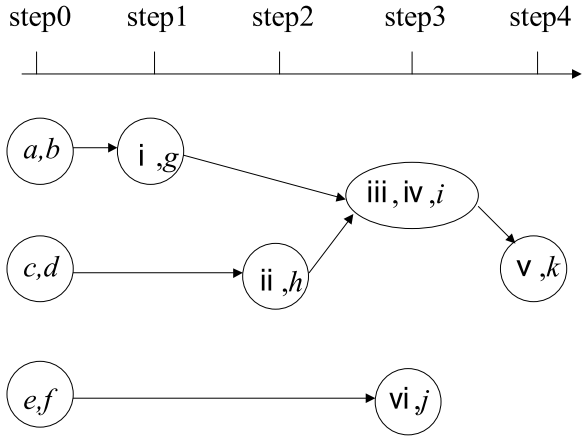

図5. 繰り返しDM分解により得られる階層化構造

2.3.クラスタリングとクラスタラベル

単語-文書行列をグラフ行列Gと考えるとDM分解により行列の対角線上の1の周辺に正方部分行列が得られる。この正方部分行列を2部グラフ形式で表すと図4のような単語と文書がエッジで結ばれる2部グラフとなる。

図6の0-1行列の対角線上に並ぶ1をすべて集めたものを行列の芯という。図6が2部グラフを与えるグラフ行列Gであるとき、この芯は2部グラフの最大マッチングと等しい。つまり図6の単語-文書行列において、ある文書のマッチング相手である単語を求めることができる。

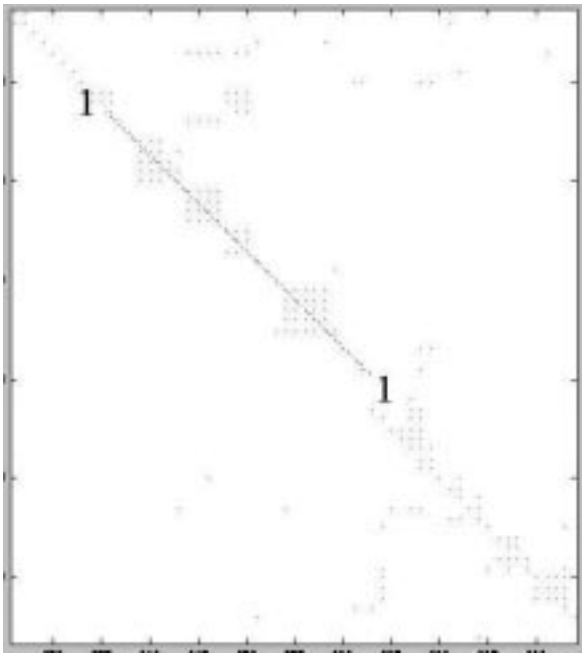

図6. 単語-文書行列をDM分解した例（横軸は文書、縦軸は単語を表す）

本手法では統合文書のマッチング相手である単語をクラスタのラベルとして扱う。

例として、図7を用いてクラスタラベルの決定方法を述べる。図7のstep0では文書a、bがクラスタリングされ統合文書ⅰが生成される。次にstep1の時点では統合文書ⅰがDM分解により文書gとクラスタリングされる。同時に統合文書ⅰの2部グラフにおけるマッチングである単語が求まる（図3(a)の太線はマッチングを示す）。この統合文書ⅰのマッチングである単語をstep0の統合文書ⅰのラベルとする（図7のL1）。

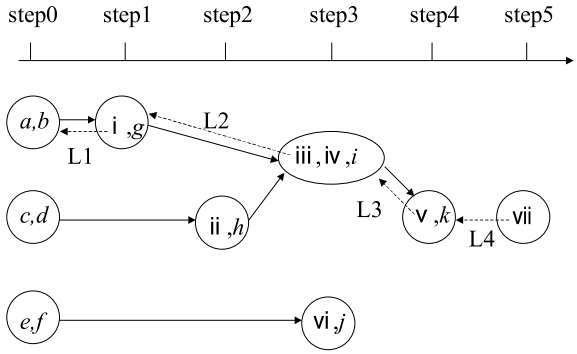

図7. クラスタラベルの決定方法

繰り返しDM分解によって最後に得られるクラスタ（図7のstep4)のラベルは同様にクラスタの統合文書（図7のⅶ)を生成し、2部グラフにおけるマッチングである単語を求めることで決定される（図7のL4）。

まとめるとクラスタラベルの獲得方法は次のようになる。

(1) 実線で示すエッジの向きは、クラスタの統合方向を表す。有向線分の始点側のノードは、有向線分の終点側のノードに統合される。
(2) クラスタが統合されたときに仮想代表文書の最大マッチングを求める。
(3) 最大マッチングから求められる単語を下位のクラスタのラベルとする。

2.4.重要文抽出による要約文生成

要約文の生成では、文書中の重要な文を抜き出す重要文抽出法[13]を用いる。これは文書中の重要語を決定し、この重要語を含む文を重要文として抽出する手法である。

本報告ではクラスタラベルを重要語と考え、これを含む文を抽出することで要約文を生成する。まとめると要約文生成は次の通りとなる。

(1) 重要文抽出法による要約文の生成を行う。
(2) この際に利用する重要語として、クラスタリング結果から得られる単語を使う。

具体的な手順としては、下記の通りとなる。

文書Dはn個の文Siから構成されているとし、その関係を次のように定義する。

$$D=\{S1, S2, \ldots, Sn\}$$

本手法では各クラスタに分類される複数文書に対して、クラスタラベルとなる単語が含まれる文Siを抽出し要約文集合Eに加えることで要約文書を生成する。

2.5.提案手法の特徴

提案手法は文書-単語間の組み合わせ構造のみに着目し、これを逐次変形することでクラスタリングを行うことを特徴とする。これと比べて他の手法では、各文書同士の類似度を計算し、その結果得られる類似度行列からクラスタリングを行う[14]。この類似度を計算する方法としては、自然言語処理の場合にはベクトル空間法を用いるのが一般的である[15]。

本手法は単語-文書行列を2部グラフ行列Gとみなしクラスタリングからクラスタラベルの決定までを一貫して2部グラフの操作によって行っている。

クラスタリング結果である個々のクラスタに対して、別途ラベル決定プロセスを介することなく、クラスタリング処理の一部としてクラスタラベルが決定されることは、本手法の大きな利点である。

自然言語処理においては、単語-文書行列のデータ表現形式により、文書の特徴を表現する。本手法においては、単語が文書中に存在しているか否かで行列の要素の値は0-1のいずれかの値を取る。またその他の方法としては、TFIDF[16]などの方法により各単語に対して重みが計算されその値を行列の要素の値とすることもある。

行列の要素の値が0-1の値を取るのか、重みを表す値を取るのかの差はあるが、単語-文書行列のデータ表現形式を取る限り、文書中で各単語の出現する順番に関する情報が保存されないなど、このデータ表現形式が持っている情報量は同じであると考えてよい。

以上のまとめとして提案手法の特徴は次の4点となる。

第1に、単語と文書の関係を単語-文書行列から2部グラフで構成し、ノードの結合関係のみを利用してクラスタリングを行う。

第2に、本手法は、単語-文書行列からDM分解を使って直接的にクラスタを生成する。これに対して従来は、各サンプルもしくは個々の文書同士の類似度を一度計算しその後、類似度の近いサンプルもしくは個々の文書同士を統合することでクラスタを生成する。

第3に、本手法は、単語-文書行列において同一クラスタに分類された文書から仮想的な代表文書を生成する。仮想的な代表文書と一緒にクラスタに文書が分類された時点で、クラスタの統合関係が成立する。

第4に、クラスタリングの開始時に得たいクラスタ数cを予め決定する必要は無く、クラスタリングの終了条件のみを決定しておけばよい。

3. 文書解析ツールREXION Pro

データをクラスタリングすることによって得られる結果は「データの要約」であると考えられ、各種知見を得るために有用に活用することができる[17]。提案手法を適用した文書解析ツールREXION Proにおいて重要語を抽出する際には、単語の共起関係[18],[19]を利用する方法により各語の評価値を導き出し、これが高い単語を重要語と判定する。これにより図8のような重要語-文書行列を内部に生成し処理を行う（pは共起関係にある重要語のペアを表し、textは文書を表す）。

	p1	p2	p3	p4
text1	1	0	1	0
text2	1	1	0	1
text3	0	1	0	1

図8. 重要語-文書行列

処理が完了しREXION Proによって解析された結果は図9に示すような各種情報可視化手法によって利用者に提示される。

図9の上（話題分布グラフ）は抽出した話題グループ（以下、話題）の大きさの分布を直感的に把握するためにチャート化したもので、例えば、最も大きな話題を選択すると対応するデータが解析結果テーブルでハイライト表示される。図9の中（解析結果テーブル）は話題をテーブル化したもので、各話題のラベル情報や要約文が格納されており、効率よく話題の内容を掴むことが出来る。図9の下（情報マップ）は関係性の高い話題をリンクして階層マップ化したもので、ツリー形状と各話題のラベル情報を俯瞰することで、どんな話題がどのように関係しているかを直感的に把握することが可能となる。

つまり、クラスタリング結果の大域的な情報を伝えることができ、階層構造のマップを見ることで複数のクラスタがどのように統合され、そのクラスタは階層構造全体でどこに位置するのかが分かる。それによりクラスタリング結果全体に対して参照すべき特徴的なクラスタがどれであるかを判断することができる（図10）。

図9. 解析結果の情報可視化例

図10. 情報マップの例

3.1. 文書クラスタリングによる相関ルール生成

相関ルールについてもっとも簡単な場合としては次のように

定義することができる。

データセット上の属性として、C_1、C_2が存在する。その属性は値としてa_i=v$_i$を取る。属性間の関係において $C_1 \wedge C_2$ がある頻度以上で成立するときに相関ルールとして $C_1 \Rightarrow C_2$ で表すことができる。通常 a_iはカテゴリ値である場合が多いが a_iが実数の値を取るときには各a_iに対して下限値と上限値を設定することで離散化できる。こうすることにより、カテゴリ値の場合と同様に相関ルールを求めることができる。このときa_iは $a_i \in [l_i, u_i]$ である。

相関ルールの具体的な例としては、スーパーマーケットで買い物をした男性の全体は574人であったとき、ビールを買った人は331人であり、また、紙おむつを買った人は213人であったとき、この二つを同時に買った人は二つの属性間の相関ルールとして $C_1 \wedge C_2$ または $C_1 \Rightarrow C_2$ の関係が成立し{ビール}->{紙おむつ}であると表現することができる。これは小売業で有名な、ビール買った人は紙おむつも同時に買う、という相関ルールに該当する。

これを集合で表すと図11となり、上記の相関ルールは集合の積の部分に相当する。

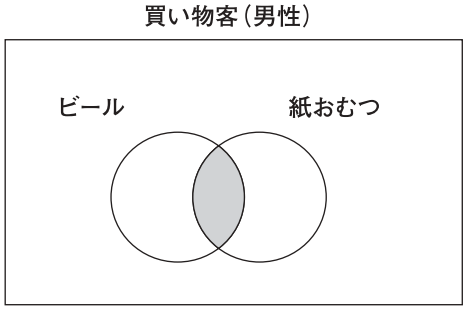

図11. 相関ルールの例

従来はこの相関ルールは属性間同士に対して適用でき、文書情報に対しては適用することはできなかった。

一方、RexionProは文書情報のクラスタリング結果としてクラスタラベルを生成することができる。このクラスタラベルは文書の内容を上手にカテゴリ値（属性値）として表したものと解釈することができる。つまり、RexionProの生成するクラスタラベルを使うことで従来は相関ルールでは直接扱うことができなかった自然言語情報をカテゴリ値の情報に変換する。これにより自然言語情報とほかの属性情報との間で相関ルールを生成することができる。

この相関ルールに関する情報可視化としては、情報マップに表示されているクラスタラベル情報を集計し、出現頻度順にチャート化した話題カテゴリレポートを用意している。
これは情報マップからツリー構造とラベル情報を読み取って生成したもので、クラスタリング結果全体を俯瞰しつつ特徴的なクラスタラベルを特定化するのに有効である(図12　左半分)。

更に、話題カテゴリレポートで着目した話題（クラスタラベル）と任意の属性情報（たとえば文書情報が記録されたときの天候）との相関を調べてチャート化した属性相関分析レポートを用意している。これは、ある話題が特定の属性情報とどのように関係しているかを把握するための機能である。例えば、クレーム記録を解析した場合など、ある不具合現象（話題）が特定の製品モデル（属性）に集中して発生している状況などを容易に把握することが出来る(図12　右半分)。

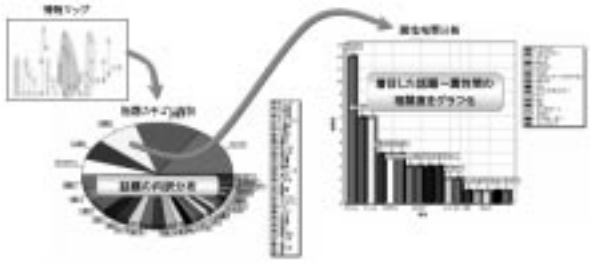

図12. 相関ルールの可視化手順

4. 適用例

複数の問題について本ツールを適用している。その概要を以下に述べる。

第1の適用例は、知識の構造化を行うことで知識基盤の構築[20]を支援するものである。領域化によって深化した知識は、非専門家はもとより専門家とっても理解することは難しい。複雑化した知識の活用のためには、知識の構造化表現が必要であると言われている。

具体的には特許文書の理解支援を目的とし、公開された特許文書を本手法により階層化クラスタリングを行った。結果の評価では発明の構成要素が概念的な階層性に基づき表現されることが分かった。これは特許文書における知識の階層化表現が実現できることを示唆する。

本論文の提案手法は文書をクラスタリングし同時に分類クラスの情報としてそのクラスのラベル情報も同時に生成する。階層的なクラスタリングと同時にクラスのラベルも同時に決定する本手法は自然言語における概念クラスタリング法と考えることも可能である。

第2の適用例は製品品質に関する複数文書を解析することで、問題点を特定化するものである。

従来は人手により品質に関する文書を閲覧し分析を行っていた。例えば科学技術振興機構（JST）では過去の事故の事例を集め、人手により分析を行いそこから教訓や知見を獲得し整備、公開する事業を行っている[21]。ここで公開される教訓や知見は専門家により精査されるので非常に正確である。しかし公開されている情報の数は多くない。

我々は国土交通省が公開している大量の自動車のリコール情報を本システムによりクラスタリングし、その結果を参照することで人手による方法と比べて効率よく知見が獲得できることを示した[22]。

本システムはクラスタリング結果である階層構造の情報と、それを構築する元となった記録文書（具体的な事例記録）をインタラクティブに参照可能である。有用な知見が得られた場合、これらの情報を一緒に提示することにより、背景情報を踏まえたより深い理解が得られると考えられる。これは知識の共有を促進する一つの方法となる。

第3の適用例は、プラント設備の運転日誌や保守記録からの知識発見への適用である。本手法でデータの要約を行うことにより、現在まで蓄積された文書情報を分析することが現実的に可能となる。

産業用機械や装置の耐用年数は一般に長い。20年から30

年に渡って継続的に使用される場合が多いが、この運転に関するノウハウの管理は言い伝えや申し送りなど人に依存する属人的な方法によって行われていることが一般的である。

本手法を、運転の経緯を記録する目的で管理されている業務日誌に対して適用した結果、「状況と対処」といったルール形式で表現できるようなクラスタリング結果を得ることができた。またノウハウの単位から元の日誌の文書に逆引きすることにより運転や保守の業務内容を大域的かつ具体的に理解することも可能となる。

具体的な手法としては、下記の通りとなる。

例えば設備保守業務では、電子化された情報として定期保全記録、突発対応記録、補修記録といった文書データが保全管理システムに蓄積されていることが多い。

こういった記録類は、概ね「**の状況のとき、◇◇した」という一対の構造を持った文書であることが典型的である(図13)。

図13. 設備保守データの例

このような形式のデータに対しては上記**(=状況記述データ)を分類キーとするようにクラスタリングを実施する。

図14. 設備保守データの知識化例

保全措置が必要とされる状況を分類のベースにすると、蓄積したデータから作業ルールを「**の状況のとき、◇◇する」という形で取り出すことが容易となる。このルールをそのまま新たな作業標準として運用することで、業務の効率化や確実性を向上させることが出来る(図14)。

また、逆に◇◇(=対処記述データ)をキーとして分類し、業務分析することも出来る。具体的には、◇◇の対処と相関の高い属性(設備名、装置名、作業種別、製造品種、等々)をつかむ事によって、◇◇作業が発生する原因やその回避策を検討することも出来る。

第4の適用例は、インシデントレポート(ヒヤリハット事例)を解析することで事故を未然に防止する、つまりリスク管理を効果的に実行するものである。

危険予知訓練といった取り組みは、労働災害リスク管理の要として既に現場に定着したといえる。この活動の本質は、個々の事例を現場にかかわるメンバー間で共有し、日常の基本動作としてリスク管理を行ってゆくという“意識高揚への取り組み”である。

しかし、このほかにもアプローチとして危険予知の題材となった事例データ(=潜在リスクのシグナル)を分析して、その傾向を把握するという直接的な取り組みも検討に値すると考えられる。

危険予知活動の記録を集約して「どのような種類の潜在リスクがどの装置に内在されているのか?」という傾向を把握する事ができれば、より少ないコストで効果的なリスク管理が実現出来ることは明らかである。大量のエラー/インシデント事例を分析することで要因を特定し事故を未然に防止するというPro-Activeな安全管理を実現することができる(図15)。

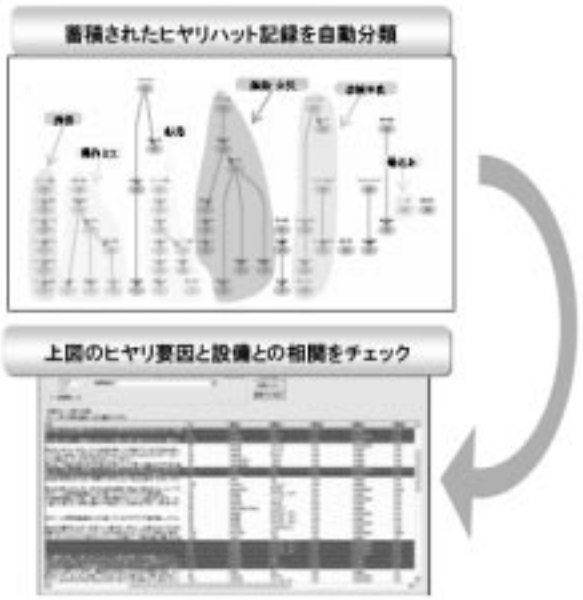

図15. ヒヤリハット記録のリスク分析例

5. おわりに

文書集合から単語-文書の関係を表す、2部グラフを生成し、この2部グラフに対して繰り返しDulmage-Mendelsohn(DM)分解法による新しい階層化クラスタリング法を提案し[23]、それを適用した文書解析ツールを開発した。

「既存の文書データ」を素材とし文書解析ツールの力を借り

ることで、今まで形式知化することが難しかった“業務の勘所”と呼ばれるノウハウを抽出し共有・伝承してゆくことがある程度可能となった。

この一連の流れのなかで解析ツールは、個別性の高い業務記録データの集合体を集約・類別することを通じてある程度一般化し、人間が俯瞰・解釈できる状態に前処理するという役割を果たしている。

つまり、あくまで解析結果を解釈しデータに内在されている情報を知識として昇華するのは、依然としてツールを操る人間側の作業であり、ツールが自動的にやるべき事柄を教示してくれるものではないということである。

しかし、逆の側面から考えると人間の持つ推論、連想、仮説構築～検証といった高度な情報処理能力を最大限に活かすためには、雑多な状態のデータをなるべく直感的に理解できる形に前処理することが重要であり、ここにこそツールを活用する意義があるといえよう(図16)。

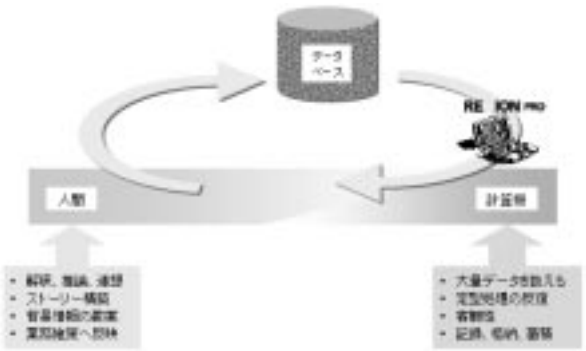

図16. 文書解析ツールと人の協調関係

このような観点で今後の展開を考えてみると、次の段階として検討すべき課題は話題間の関係性を把握しようとする試みではないだろうか?

要素還元的思考法(部分の集合が全体であるとする考え方)だけでは見えてこない仮説を明らかにするためには、関係性への着目がキーとなるだろう。

弊社では本件について既に幾つかの取り組みに着手しているが、これについては後日機会を改めて報告を行いたい。

参考文献

1. 岩崎　学:「テキストマイニング」ー特集号によせてー,電気論C, 125,5,pp. 681(2005-5)
2. 保田明夫:テキスト・マイニングの概要,電気論C, 125,5,pp. 682-689(2005-5)
3. 佐藤洋一,尾関和彦:単語間意味関係のグラフ理論的解析,電子情報通信学会技術研究報告NLC90-52 (1991)
4. 山下達雄,松本裕治:コスト最小法と確率モデルの統合による形態素解析,情報処理学会研究報告 97-NL-119 (1997)
5. 松川智義,中村順一,長尾真:共起関係に注目したDM分解と確率的推定による単語のクラスタリング,情報処理学会自然言語処理研究会報告72-8 (1989)
6. Dulmage, A. L., Mendelsohn, N.S.: Covering of bipartite graph, Canad. Jour. Math., 10, pp.517-534 (1958)
7. Dulmage, A. L., Mendelsohn, N.S.: A structure theory of bipartite graphs of finite exterior dimension, Trans. Roy. Soc. Canad., Sec. III, 53, pp.1-18 (1959)
8. 藤重　悟,グラフ・ネットワーク・組合せ論,工系数学講座,18, 共立出版,(2002)
9. 佐藤洋一,尾関和彦:単語間意味関係のグラフ理論的解析, 電子情報通信学会技術研究報告NLC90-52 (1991)
10. Murakami, E., Terano, T.: Information Clipping from Internet Documents with Similar Contexts, IEICE Technical Report,Vol.103,No.306,AI2003-60 (2003)
11. Guha, S., Rastogi, R., Shim, K.: CURE: An Efficient Clustering Algorithm for Large Database, Proc. ACM SIGMOD Int'l Conf. Management of Data, ACM Press, New York, pp. 73-84 (1998)
12. Han,J., Kamber,M.: Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers (2001)
13. 佐藤理史,奥村学:電脳文書要約術-計算機はいかにしてテキストを要約するか-,情報処理,Vol.40,No.2,pp.157-161 (1999)
14. Hastie, T.,Tibshirani, R.,Friedman, J.: The Elements of Statistical Learning Data Mining,Inference, and Prediction, Springer (2001)
15. Salton, G.,Lesk,M.,E.: Computer evaluation of indexing and text processing, Journal of the ACM, 15(1),pp. 8-36 (1968)
16. Church, K.W., Gale, W.A.: Inverse document frquency(IDF): A measure of deviation from Poisson, Proc. Of the Third Workshop on Very Large Corpora, pp.121-130 (1995)
17. 神嶌 敏弘,データマイニング分野のクラスタリング手法(1), 人工知能学会誌,Vol.18,No.1,pp.59-65 (2003)
18. 原　正己,中島　浩之,木谷　強,単語共起と語の部分一致を利用したキーワード抽出法の検討,情報学研報告,NL106,pp.1-6 (1995)
19. 大澤幸生,ネルス　E. ベンソン,谷内田正彦:Key Graph: 語の共起グラフの分割・統合によるキーワード抽出,電子情報通信学会論文誌,Vol.J82-D-I,No.2,　pp.391-400 (1999)
20. 松本 洋一郎, 学術創成研究費プロジェクト「学術創成のための知識の構造化とネットワーク型知識基盤の構築」, http://www.t.u-tokyo.ac.jp/archives/2004/0903.html,2004
21. 失敗知識データベース:http://shippai.jst.go.jp/fkd/Search
22. 村上英治,越水重臣, 林純也, 寺野隆雄:テキストマイニングを使ったリコール情報からの問題点抽出とナレッジマネジメントへの適用,経営情報学会2005年春季全国大会,pp.246-249 (2005)
23. Murakami,E., Terano,T.: Multiple Document Summarization and Visualization through a Combinatorial and Hierarchical Clustering Method, 16th European Conference on Artificial Intelligence, The 1st European Workshop on Chance Discovery, ECAI 2004, pp.194-203 (2004)

著者所属

村上　英治　　アドバンスオートメーションカンパニー
ソリューションマーケティング部

木幡　真望　　アドバンスオートメーションカンパニー
ソリューションマーケティング部