技術文書群からの技術・組織動向の抽出と可視化による戦略的な技術管理

Strategic Management of Technology by PhyloSystematic Technology-Organization Visualization

近年，相次いで特許などの技術文書から最新の技術動向を抽出するテキストマイニング技術が開発されている。本報告では，当社が開発した技術文書群から技術・組織の動向を抽出し，企業の戦略策定と研究開発業務の効率向上を支援するソリューションの開発について述べる。本ソリューションでは，生物系統学と情報知識学での系統構造抽出アルゴリズムを応用し，系統図による技術の生成・発展過程や組織内の共同執筆関係の分析を実現可能とした。技術系統の中心にある基本技術と枝にある先端技術や，社内外の組織が採用している技術を参照し，市場価値の高い技術を採用するための参考とすることができる。企業での新技術管理や部署・企業間の比較・連携支援での活用事例を説明する。

高梨　勝敏　Takanashi Katsutoshi
塚原　朋哉　Tsukahara Tomoya
佐藤　俊也　Sato Shunya

１．はじめに

事業を継続・成功させるためには，品質と生産性を向上させるとともに，市場や顧客ニーズの変化に対応できる新事業を効率的に創出することが重要である。技術に対する経営の知識体系は MOT[1]（技術経営）により整理された一方，収集・蓄積された顧客や技術の情報を分析し，全体の傾向把握や新たな事実の発見を効率化するためのソフトウェア技術も発達してきた。大量の数値データを分析するデータマイニングに加え，テキストデータを定量的に扱うテキストマイニングの普及が進んでいる[1)2)]。

本報告では，従来のテキストマイニング手法に加えて，情報の系統的な発生関係を自動抽出する手法を用いて新規開発したソリューションについて述べる。この手法により，企業の研究開発管理情報や特許情報から技術・組織の変遷を抽出し，経営戦略策定および研究開発計画時の根拠となる情報提供を実現する。

２．背景

当社をはじめとしてテキストマイニングソリューションを提供する企業では，業務に効率的・効果的に導入するために業務別ソリューションやテンプレートを提供している。対象業務は，市場から得られる情報の分析（顧客の声分析など），企業が保持する技術や人材の分析（特許分析など），製品の品質に関する情報の分析（不具合分析など）に分けられる。それぞれの業務適用事例をこれまで報告してきた[3)4)]が，本開発では技術・人材の分析を

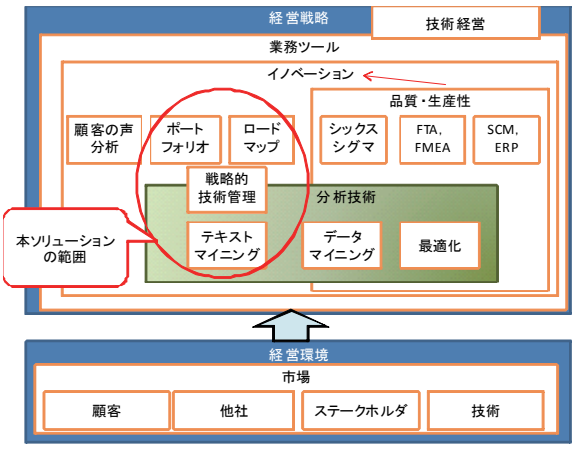

図１ 技術経営でのテキストマイニングの位置づけ

[1] Management of Technology

より高度化し，ポートフォリオやロードマップなどの経営上の活動と連動して戦略的に技術管理をおこなうため，新しいテキストマイニング技術を適用した。(図 1)。

本開発の背景には，企業での事業戦略部門が経営者の戦略策定と研究開発部門の業務を支援する際に，研究計画・報告書，特許および文献から現在の市場の状況を分析し，意思決定の根拠となる適切な情報をタイムリーに提供できるようにする狙いがあった。従来の戦略策定と研究開発計画時には，調査をするために自分の知っている範囲以外の情報を収集する効率が悪かったり，会議の際に必要な情報が揃っていない場合には仮定に基づいて議論をする必要があった。そこで，計画策定時，実行後の見直し時および次期計画策定時に自社・他社・市場の情報を容易に得られるようにして PDCA サイクルを支援するソリューションが求められていた（図 2)。

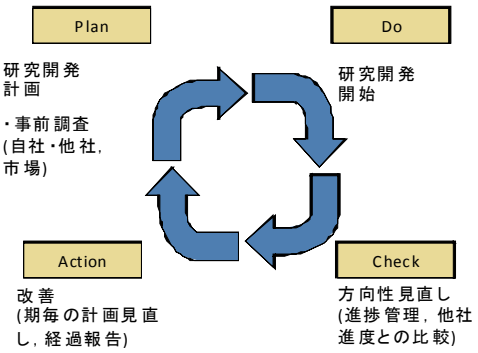

図 2 戦略的技術管理の PDCA サイクル

3．目的・目標

企業の戦略部門での構想に基づき，PDCA サイクル実行の効率向上を目的として，ソリューション開発の目標を設定した。

(1) 着目している分野の情報を抽出できるだけでは知らない情報を得るための効率が悪いため，当該分野の全体を俯瞰でき，適宜絞り込んで文書に到達可能とする。
(2) 現在メインとなっている技術を抽出できるだけでは新技術の動向を追うことができないため，技術と組織が生成・発展してきた過程を知ることを可能とする。

(1)では，計画の策定や見直し時に自分が知っている範囲で判断してしまうことを防ぎ，自社内での研究計画・報告書や，社内外の特許および文献から，事業をとりまく市場の情報を得たうえで意思決定できるようにする。(2)では，基本技術と先端技術を知った上で，自分の計画が市場に対しどのような価値を提供するか判断できるようにする。

4．狙い

テキストマイニングは情報の抽出と可視化で構成される。それぞれについて，以下の技術開発の狙いを定めた。

4.1 情報の抽出

文書から技術用語を抽出する従来の方式は，大きく以下に分けられる。

(1) 浅い意味解析
各文書に出現する単語の出現数から，単語の出現数の偏りや，各単語が文書・パラグラフおよび文に同時に出現する確率を求め，特徴的な出現傾向を示す単語の組を抽出する [5]。
(2) 深い意味解析
単語間の係り受け関係を解析し，主語-述語-目的語など，文書の主題となる単語の組を抽出する [6]。

本開発の目標は全体俯瞰と系統の抽出であるため，(2)の深い意味解析の高度化を追及するよりは，(1)で統計的

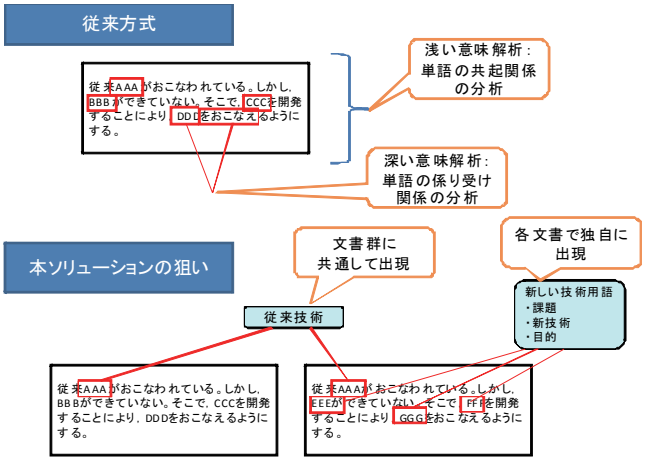

図 3 技術文書分析でのテキストマイニングの狙い

解析手法を高度化するほうが効果が大きいと考えた。そこで，文書群から従来技術と新技術の用語を抽出する際に，文書群に共通して出現する用語を従来技術として抽出するとともに，各文書で独自に出現する用語を新技術として抽出できるように，浅い意味解析の統計的手法を高度化することを狙いとした（図 3)。

4.2 情報の可視化

抽出した情報を可視化する技術は，大きく以下に分けられる。

(1) 階層構造の可視化
技術用語間の分類関係（大分類から小分類へ）や，構造（部品構成など）といった，企業が持つ技術の知識体系を可視化する [7]。

(2) ネットワーク構造の可視化
単語の共起関係（文書，段落や文で同時に出現する確率）など，文書群での単語の出現傾向を可視化する [8]。

(1)は組織で共有できる知識体系を構築するのが目的であり，情報の動向分析とは異なる分野の技術である。(2)はネットワーク構造を対象としており，技術の生成発展での階層構造を対象とするものではない。そこで，上記目標（情報の俯瞰と絞り込み，技術・組織の生成発展過程の可視化）に到達するために，情報の階層構造とネットワーク構造の両方を抽出・可視化できるテキストマイニング技術を新規開発することを狙いとした。

５．モデル構築

5.1 モデル構築の背景

文書群から用語の系統関係を抽出する手法を開発するにあたり，従来手法の調査をおこなった（図 4)。情報知識学の分野では，系統関係が知られていない文書群から系統を類推する際にマイニング手法が用いられている。例えば，古典の文書で書写により複数の版がある場合に，単語の出現数や章構成の類似性から原本を類推することに応用されている [9]。

また， DNA 解析技術の発展にともない，生物系統学ではデータマイニングを用いた分析手法が用いられている。これにより，生物の形態の類似性から系統を予想する手法では得られなかった DNA の類似関係の客観的な情報を得ることができる。

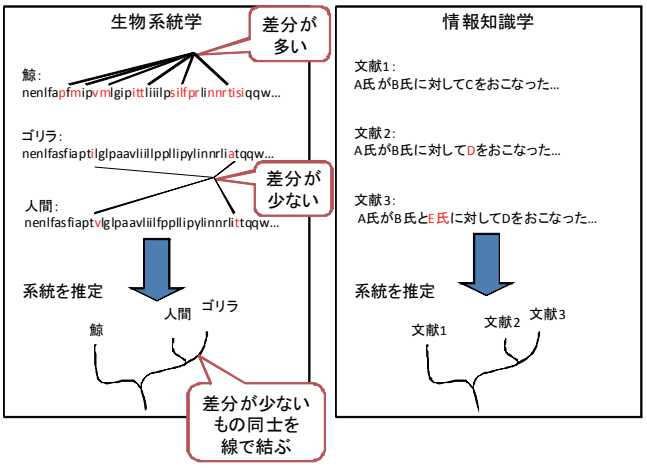

図 4 生物系統学と情報知識学の関係

当社は，これら手法を参考にして技術文書からの系統情報抽出・可視化技術を開発した。

5.2 モデルの内容

系統情報抽出・可視化の流れを図 5 に示す。従来のテキストマイニングでは，文書群から単語を抽出し，単語の出現数から各単語の重要度の計算と，単語間の関連度の計算をおこなう [4]。これにより，単語同士が関連度で結ばれたネットワーク構造を生成する。系統図生成では単語間の関連度を入力として，関連度がより高いもの同士の関連を残して木構造となるように関連を削除していく。

系統の生成手法には複数のアルゴリズムが存在する [10]。系統の木構造は枝とノードで構成されるが，1 個のノードが複数の親を持つことを許すアルゴリズムと，許さないものがある。前者の場合は，ネットワーク構造を含んだ木構造となる。本ソリューションでは，表示を見やすくするためにネットワーク構造を含まない木構造とすることにした。これにより，各ノードが持つ親は 1 個のみとなり，複数の従来技術からの新技術の生成を可視化できない。そこで，ネットワーク構造の中で関連度の高いものは木構造と異なる関連線で表示するなど，画面表示で工夫する方針とした。

系統生成アルゴリズムのひとつである，プリム法を用いた。処理の流れを以下に示す。

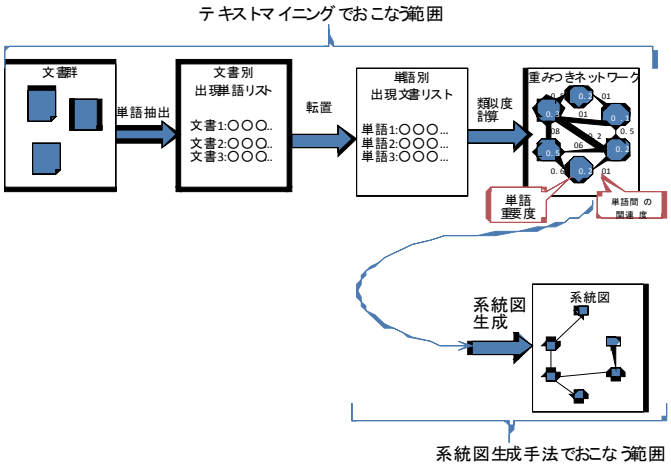

図 5 系統図生成の流れ

(1) 単語 1 個を任意に選ぶ。これを a とする。
(2) a と最も関連度の高い単語を 1 個選ぶ。これを b とする。a-b 間の枝を表示する。
(3) (1)～(2)で選んだ単語群を G とし，G と最も関連度の高い単語を 1 個選ぶ。これを c とする。G-c 間の枝を表示する。
(4) すべての単語が選ばれるまで(3)を繰り返す。

関連度を枝の長さ(関連度が高いほど短い)とすると，このアルゴリズムは枝の長さの合計が最小となるように枝を選ぶことに相当する。通常の最短経路問題（巡回セールスマン問題）では枝が一筆書きとなるように選ぶ際に計算量が O(n!)のオーダとなり処理時間が問題となるが，系統図生成では単語のグループ G から最短となる枝を選べばよいので，計算量は $O(n^2)$のオーダで済む。

なお，(1)ではどの単語を 1 個選んでも同じ木構造となるが，本ソリューションでは重要度が最大の単語を用いた。当該単語を最上位の親として各単語を配置することにより，文書群の中で重要な技術を中心とした技術の系統を視覚化することができる。

また，技術用語の系統図を生成するために単語間関連度を用いているが，図 5 での転置をおこなわずに文書間関連度を求めることにより，文書の系統図を生成することもできる。これは図 4 の生物系統学で生物を文書，DNA を単語に対応して考えると，生物系統の生成に相当するのは文書系統のほうである。技術の系統を抽出する狙いにより，単語間関連度を用いた手法としている。

６．評価

マイニングシステムを利用する際には，系統生成結果を用いて技術用語の系統図を単語マップの形で画面に表示する。画面例を図 6 に示す。画面左側に文書一覧を表示し，系統図で単語を選択することにより，文書を絞り込むことができる。以下に，ソリューション開発の目標に対する達成度評価を説明する。

6.1 全体の俯瞰と絞り込み

図 6 は，液晶に関する特許の分析例である。2006 年～2007 年に出願された特許の中で液晶の表示技術に関する特許 6,565 件を用い，課題欄に形容詞「暗い」の係り受けが出現するもの 108 件を分析対象としている。課題欄，手段欄およびすべての内容から単語を抽出し，それぞれの系統図を表示している。前章で述べた系統図生成アルゴリズムにより生成した木構造を実線で表示し，木構造以外で関連度の高い単語間を点線で表示している。これら単語と線の表示により，特許群全体で技術用語がどのように分布しているかを知ることができる。マップ中の単語をマウスで選択すると，当該単語が出現する文書および，文書に出現する単語を絞り込んで表示することができる。これにより，全体を俯瞰しつつ絞り込んで目的の文書に到達するという目標を実現している。

6.2 技術と組織の生成・発展過程の可視化

図 6 の手段欄では「バックライト」を木の中心として，各技術用語を配置している。バックライトの入射面や裏面などが共通の技術であることを示唆している。ソース-ドレーンや光学フィルタは，各特許に特徴的な技術であることを示唆している。

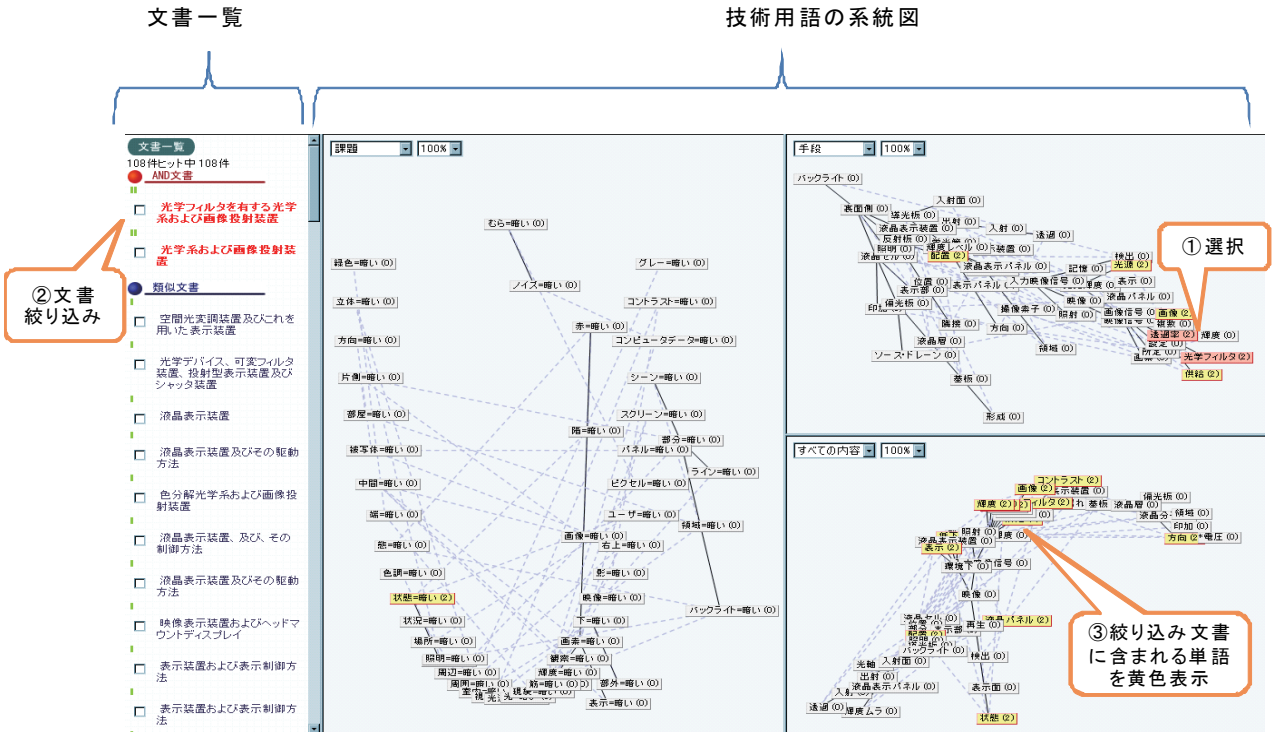

図 6 マップ画面の例

利用者は，手段のエリアで「光学フィルタ」と「透過率」を選択している。これら単語は赤色で表示し，文書一覧エリアでは当該単語を含む 2 文書に絞込み，「AND 文書」の一覧に表示している。当該 2 文書との関連度の高い順に，他の文書を「類似文書」の一覧に表示している。文書の内容を参照することにより，光学フィルタと透過率に関する特許の内容を確認することができる。また，AND 文書に含まれる単語を黄色で表示している。課題欄では「状態=暗い」，手法欄では「光源」「配置」など，すべての内容欄では「コントラスト」「方向」などを表示している。これにより，光学フィルタの透過率について，画面が暗くなる課題に対する配置や方向の工夫による解決手段があることを示唆している。これら単語を含む文書の本文を参照することにより具体的な解決手段を確認し，示唆された技術の生成・発展過程を検証することができる。

系統図の精度を評価するために，各特許文書の「手段」欄を参照して，論旨の中に登場する単語を確認した。系統図との比較結果を表 1 に示す。例えば，「手段」欄で液晶層の基板を形成することについて論旨が展開されている場合は，液晶層−基板−形成という単語の組が使用されているとみなし，系統図でこれら単語が関連線で結ばれているか確認した。全文書中 79%の文書で，系統図の木構造の各部分に一致する論旨が存在した。木構造と一致していない論旨が 16%の文書に存在したが，「配置」「入射・出射」「光源」など，液晶関連の特許に多く出現する単語が係り受けに入っているため，系統図中の広い部分と関連していることが原因と考えられる。これらの関連は，ネットワーク構造をあらわす点線によって確認することができる。

また，系統図に現れない単語を用いた論旨が 6%の文書に存在した。これら単語は重要度が低いために表示対象外となったものである。これら単語は，対象文書を特定分野・期間または特定組織の文書に絞り込むことにより，重要度が上がり分析対象とすることができる。

7．業務への適用

研究管理業務への応用例を図 7 に示す。研究計画・報告および特許を管理する既存のシステムが存在する。当該システムに蓄積された文書群から，技術用語と組織情報を抽出して可視化する，Know-Who 検索システムを開発した。技術用語の系統や，組織・人の協業関係を俯瞰し，必要な情報に絞り込むことが可能である。当該シス

表 1 系統図と特許本文中の論旨の比較

系統図との比較	「手段」欄の論旨で使用されている単語	文書数
木構造の一部と一致	(液晶層 or 基板)-形成	19
	(バックライト or 導光板 or 偏光板 or 液晶パネル or 光源)-(入射 or 出射)	13
	液晶セル-印加	11
	(画像 or 透過率)-輝度	7
	(液晶パネル or 液晶層 or 照明)-照射	6
	光源-検出	3
	(画像信号 or 映像信号)-輝度-検出	2
	表示パネル-輝度レベル	2
	(照明 or 反射板)-位置-隣接	2
	表示-検出-記憶	2
	入力映像信号-記憶-検出	2
	透過率-光学フィルタ	2
	撮像素子-領域	2
	(バックライト or 液晶表示装置)-蛍光管	2
	(表示パネル or 撮像素子)-配置	2
	バックライト-液晶セル	1
	偏光板-液晶層	1
	撮像素子-配置	1
	画像信号-設定	1
	計	81 (79%)
木構造と不一致(単語間が枝 3 本以上離れている。)	配置-表示-光源	5
	入射-(基板-配置 or 透過率 or 検出)	3
	入射-出射-光学フィルタ	1
	光源-出射-画素	1
	位置-映像-照射	1
	液晶層-画素-輝度	1
	撮像素子-輝度	1
	光源-輝度	1
	光源-位置	1
	照明-光源	1
	計	16 (16%)
未抽出(※で示す単語の重要度が低いため)	温度※-検出	1
	音声※-記憶	1
	(ランプ※or 電極※or 遮光※or レンズ※)-配置	4
	計	6 (6%)

テムは,系統可視化に加えて以下の機能を実装している。

(1) ヒートマップによる定量情報の視覚化

研究予算など，定量情報を単語ごとに集計し，単語の背景に色付きの円で表示している。円の半径が金額などの定量情報を表し，円の色は単語の出現頻度にもとづく重要度を表すなど，系統図中の定量情報の分布もあわせて俯瞰できるようにした。

(2) 日付情報による期間分析

研究開始・終了などの各イベントの年月日情報を用いて，期間に関する分析機能を追加した。期間を指定した文書の絞込みや，異なる期間で系統図を比較することにより技術や組織がどのように変化したかを知ることができるようにした。

(3) 分析対象文書群の更新への対応

既存システムの文書や研究情報を格納したリレーショナルデータベースを巡回し，テキストデータを収集している。系統図生成の入力情報の更新をタイムリーに表示に反映させることができる。

また,利用者が着目した単語を画面の中心に配置して,関連する単語を関連度にもとづき同心円状に配置することにより単語間の関連の大小を分かりやすくする工夫や,単語を時系列に配置して時間的変遷を分かりやすくする工夫もおこなっている（図 8)。本システムは試行運用を開始し，今まで得られなかった関連情報を得られることで情報利用が活性化している。

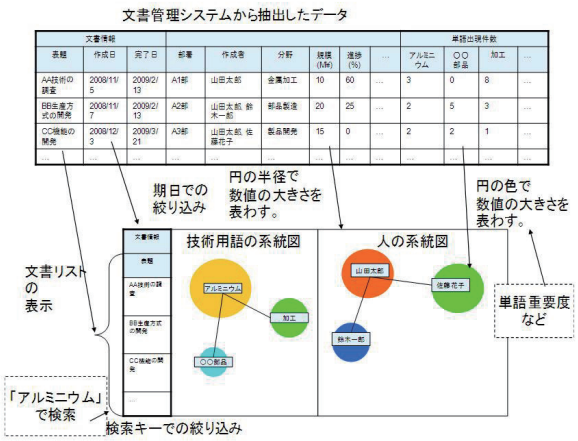

図 7 Know-Who 検索システムの模式図

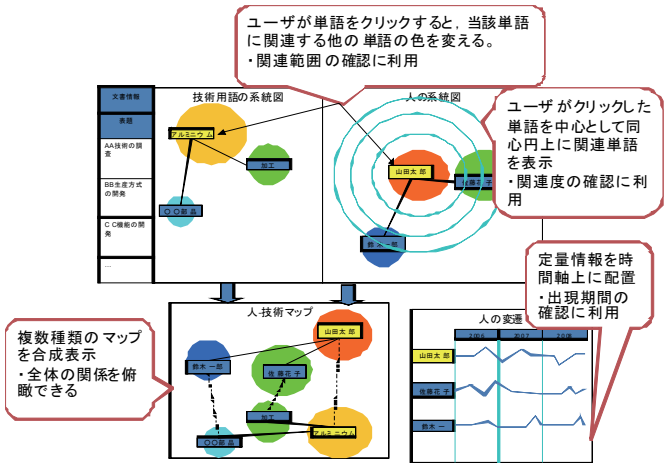

図 8 ユーザインタフェースの応用例

8．今後の課題

今回開発した系統抽出・可視化技術では木構造生成アルゴリズムを用い，各単語が親を 1 個のみ持つ仕様とした。これにより，系統図の複雑さを低減することができるが，親子関係の候補となる 2 個以上の関連を表現できない。現状では木構造以外で関連度の高いものは点線で表示することにより対処しているが，今後は木構造に加えてネットワーク構造の抽出にも対応したアルゴリズムを採用するなど，改善を進めていく。

9．謝辞

本ソリューションの開発は，日立建機株式会社殿のプロジェクトとして実施した。開発では同社の多大なご指導とご協力をいただいている。関係者各位に感謝の意を表する。

10．おわりに

経営戦略の策定や研究開発の計画時に市場調査を効率化するソリューションを開発した。生物系統学を応用した系統図生成による全体把握と，絞込みによる目的文書への到達を効率化する仕組みを考案した。従来のテキストマイニングソリューションに比べて，より業務に適用しやすいシステムを実現することができる。プロジェクトの定性情報の管理効率を向上させる，技術経営を実践する等，応用範囲を広げていく。

参考文献

1) 那須川 哲哉：テキストマイニングを使う技術/作る技術 - 基礎技術と適用事例から導く本質と活用法，東京電機大学出版局（2006）
2) 那須川 哲哉:テキストマイニングの普及に向けて - 研究を実用化につなぐ課題への取組み - ，人工知能学会誌，Vol.24, No.2, pp.275-282（2009）
3) 塚原 朋哉 他：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) 塚原 朋哉 他:複数視点からのテキストマイニングによる設計品質の向上，日立 TO 技報 第 13 号，pp.54-59（2008）
5) 山本 優樹 他：共起語ネットワーク特徴の言語・文書非依存性に基づくキーワード抽出と見出し語の予測による性能評価，人口知能学会論文誌 Vol.24, No.3, pp.303-312（2009）
6) 西山 莉紗 他:未来技術動向予測のための技術文書マイニング, 第 21 回人工知能学会全国大会予稿集, No. 2H5-3 (2007)
7) 今村 誠 他:技術文書からの用語知識の自動獲得方式の検討, 情報処理学会 研究報告 – デジタルドキュメント，Vol.2007, No.34, pp.25-32
8) 前野 義晴 他：ネットワークにおけるノード発見，人工知能学会論文誌，Vol.24，No.5，pp.376-385（2009）
9) 矢野 環：古典籍からの情報発掘 –再生としての生命誌, ネットワーク-, 情報知識学会誌 Vol.17, No.4, pp.235-242 （2007）
10) R. Graham and P. Hell : On the history of the minimum spanning tree problem, Annals of the History of Computing Vol.7, Number.1, pp.43-57 (1985)

高梨　勝敏　　1995 年入社
ナレッジソリューショングループ
知識管理システムの開発，CoreExplorer の開発，知識処理ツールの研究開発
takana@hitachi-to.co.jp

塚原　朋哉　　1997 年入社
ナレッジソリューショングループ
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グループ
CoreExplorer，テキストマイニングツール，知識管理システムの拡販，コンサルティング
shu_sato@hitachi-to.co.jp