科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　２４　日現在

機関番号：　１２４０１
研究種目：　基盤研究(C)
研究期間：　2010～2013
課題番号：　２２５３０２６１

研究課題名（和文）特許公報等のテキストマイニングによる「選択と集中」戦略の立案に関する研究

研究課題名（英文）Study on Selection and Concentration Based on Text Mining of Patent Publication

研究代表者
菰田　文男（KOMODA, Fumio）

埼玉大学・経済学部・教授

研究者番号：６０１１６７２０

交付決定額（研究期間全体）：（直接経費）　　3,000,000円、（間接経費）　　900,000円

研究成果の概要（和文）：　本研究の目的は、日本企業の「的確な選択と集中」に資するテキストマイニング手法を提示することにある。一般にテキストマイニングの難しさは、重要な単語が、膨大なテキストデータの中に埋もれており、発見が難しいことに起因する。この困難を克服するために、二つの手法を提示した。第一に、デジタルデータ上で解析対象テキストにアノテーションを付与することによって、マイニングの精度を高める手法である。第二に、単語セットを作成し、それを的確に進化させるという手法である。この二つにより作成された共起行列に多変量解析やネットワーク分析を適用することによって知識発見が可能になることを多くの事例で論証し、発表した。

研究成果の概要（英文）：The aim of this study is located in discovering the technique of text mining, which contributes to planning valid "selection and concentration" strategy of Japanese companies. In general, studies of text mining technique face to difficulties that a few important or critical words are hidden in a huge quantity of noise or meaningless words. Our study proposed two ideas or techniques in order to solve this difficulty. First, in order to make mining more accuate, original text data are given a variety of value added in digital space by attaching annotation such as words set, underline, linking origina text data with another data and so on. Second, word set are evolved in digital space or relational database into the unmistaken direction. Finally, multivariate analysis and network analysis are applied to this coincidence matrix. Our technique is expected to discover a lot of knowledge, which will play important role in planning "selection and concentration" strategy in Japan.

研究分野：　社会科学

科研費の分科・細目：　経済学,経済政策

キーワード：　選択と集中　日本企業　技術経営　テキストマイニング　テータマイニング　社内知識共有　データベース　特許公報

１．研究開始当初の背景

日本企業の低迷の理由が議論されるようになって久しい。しかし、少し遡って長い時間軸で振り返って見ると、日本企業はさまざまな困難をその都度解決し成長してきたのである。戦後の乏しい資金や外貨を重点部門に傾斜配分して効率的に利用し、海外からの技術導入と自力の研究開発などの企業努力を積み重ね、競争力を少しずつ強めてきた。1970 年代の変動相場制度移行による円高や石油価格の高騰も産業ロボットや省エネ技術開発によって乗り切り、その後のアメリカとの貿易摩擦問題も産業構造の高度化によって克服してきた。

しかし、1991 年にバブル経済が弾けて以後、日本企業は次第に国際競争力を失い収益は低迷する。困難に対する的確な対応が出来ないまま「失われた 20 年」とも言われ、現在に至っている。

今後、日本企業が新たな環境に適応して国際競争力を回復するためには、有望な事業分野、研究開発分野を的確な選び出し、そこに重点的に経営資源を集中的に投入するという事業構造の改革が必要になる。

しかしそのためには、他の何よりも将来の有望な事業、顧客の嗜好やニーズ、世界の技術開発のトレンドなどを正しく読み取ることができる「目利き力」が必要となる。優れた目利き人材の存在こそ、今後の日本企業の盛衰を決定づける最重要要因の一つである。

ところで、従来の目利き人材の資質は、彼らの個人的な経験、知識、勘など、属人的な資質に依存していた。しかし、経済システムが、重厚長大型産業から加工組み立て産業にシフトし、ものつくりがサービスや情報にシフトし、技術の融合化・複合化が進み、市場や生産の立脚点の中心が国内から海外にシフト(グローバル化)するとことによって、目利きに求められる知識の量は飛躍的に多くなり、しかも多くの異質な事業・技術・ニーズ・国の市場にまたがる知識が求められるようになり、しかもそれが不断に激し変化するようになってくる。したがって、従来のような個人の経験や知識だけでは的確な判断・意志決定が出来なくなってきているのである。

このような現実を背景として、さまざまなデータを利用し、そこから意味を発見するという研究が進んできた。データマイニングと言われるこのような研究は、複数の商品の「併売」から消費者の購買動機を発見するなどの多くの研究成果が得られている。

本研究は以上のような経済システムの変化と研究の流れの中に位置づけられている。

２．研究の目的

データマイニングの手法を企業の経営分析や、さらに広く社会現象に適用し、新たな知見や意味を発見するという研究は過去にも多く見られる。しかし、これらの多くの先行研究が企業経営という実践の場での利用という観点で見ると十分な成果を上げないままにとどまっていることも否定できない事実である。その理由は、データマイニング研究に用いられるリレーショナルデータベースとして構造化されたデータ(数値データ、文字データなど)だけでは、真に必要な豊かな意味や知見を発見するには不十分だからである。リレーショナルデータベースの各種の「属性」に合う形で、本来のデータが加工され、その中に入り込まないデータはそこから切り捨てられてしまうのである。リレーショナルデータベースとして構造化される前の段階の、自然言語で叙述されたテキストデータの中身にこそ、真に価値ある意味や知見が含まれているのであり、したがって構造化されていないテキストデータを解析して意味を発見することが現在の目利き力の獲得にとって不可欠となっているのである。

しかし、誰もが認めるように、自然言語をそのまま統計解析することは容易ではない。自然言語の中から形態素を抽出し、出現頻度情報を得ることなどは、最近のテキストマイニングツールの目覚ましい進歩によってかなり正確に得られるようになっていることは事実である。しかし、たとえば企業の戦略を構築するために目利き人材が知る必要がある単語(たとえば将来を指し示すような単語など)は、現時点のテキストデータの中での出現頻度が少なく潜在化されているので、単に形態素として発見し、出現頻度の情報を得たとしても、それが将来を指し示す単語であるかどうか簡単には理解できない。しかも、膨大な形態素の中に埋もれているので、発見自体がきわめて難しいというのが現実である。

さらに人間の脳が持つ創造性や新しい知識の創出の仕組みは未だほとんど解明されていないが、少なくとも確実にいえることは単にテキストデータを読んだり、単語の出現回数を知るというだけでは新しい知識の創造は不可能な場合が多い。人間の発想や新しい知識の獲得は、各個人の脳の中にストックされている過去の知識の体系の中に位置づけられてこそ、はじめて可能になるのであり、形式的に関連づけ、接ぎ木するだけでは真に価値ある知識の獲得が可能になることはない。しかも、この脳の中にストックされている知識の体系は、各個人ごとに異なっていて、各自に必要な形で固有に保存されている。コンピュータシステムとの類似性で言えば「データ圧縮」されて保存されているが、この圧縮のされ方が個人ごとに異なっていると言える。単に複数の人間が議論したり意見を交換しても、互いに十分な理解に達しない場合が多い理由の一つは、ここにある。このような事実は、知識とは暗黙知であることが多く、形式知化あるいは言語化されにくいという指摘として論じられている。

複数の人間が直接議論しても、了解出来るようになるためには相手から得られた知識を、自分自身の知識体系の中に的確に位置づけ直すように、成形し加工することか必要であるが、この作業は容易ではないのであり、このことがコミュニケーションの困難さや誤解の発生理由なのである。ましてや他人が文字として書いたテキストを読んだだけでは十分な理解に到達できず､自分とって価値のある知識として組み替えることはますます難しくなる。過去のテキストマイニング研究の大部分がそうであるように、他人によって書かれたテキストをそのまま解析しても、解析者がその真意を読み取ることは難しい場合が多い原因はここにあると言える。

このような限界を克服してテキストデータから、企業の意思決定や研究開発戦略の立案のように、未来予測という曖昧で複雑な知識を発見するためには、オリジナルなテキストデータを自分の頭の中に蓄積されている自身に固有な(言い換えれば普遍性の無い)知識体系の中に位置づけることができるように、付加価値を与え成形しておくことが必要である。言い変えれば、知識を自分に合う形に「構造化」しておくことが必要なのである。このような「知識の構造化」については、人工知能研究、工学教育、失敗学等のさまざまな観点から試みられており、成果を上げてきているが、このような研究をさらに進めることが求められているのである。

そのための今後の試みとして重要な一候補になるのが、デジタル空間において、さまざまなアノテーションを付加するという手法である。このような研究は多くの箇所で取り組みが始まっているとはいえ、未だ端緒段階であり今後の研究が必要になっている。

このような形で自身の知識体系に合う形でオリジナルなテキストデータがデジタル空間上で成形・構造化した後で、それを対象としてテキストマイニングツールを適用すれば従来のテキストマイニング研究では得られなかった重要な知見が得られると期待される。

本研究ではさまざまな付加価値としてのアノテーションが提示されるが、そのなかでもとくに重視されるのが「単語セット」である。単語セットという視点からオリジナルテキストデータを構造化すれば、それを導きの糸としてテキストデータの統計解析が容易になり、またそれから得られる知識も豊かになる。したがって、多次元尺度法、クラスター分析などの多変量解析、ネットワーク分析などの解析手法を試みて、最も適したテキストマイニング手法を獲得することが可能になる

以上のように、本研究の目的は、「デジタルネットワーク上に「書き込み空間を持つ知識創出・共有システム」を構築し、そこでアノテーションを付与するなどの構造化処理を施すことによって、テキストマイニングの精度を高めることにある。

３．研究の方法

以上から理解されるように、本研究は二つの柱から成っている。第一に、デジタルネットワーク上にデータベースを構築し、それにテキストデータをインポートし、そのデータにさまざまなアノテーションを付加することによって、意味の発見を容易にするための研究であり、第二に､そのデータを統計解析することによって意味を発見する研究である。

第一の柱としては、Salesforce 社の提供する商用クラウドサービスを利用し、その中に特許公報、科学技術にかんする学術論文、企業のプレスリリースなどをインポートし、それにさまざまなアノテーションを付与できるシステムを試作した。アノテーションとしては、「コメント」「アンダーライン」「単語セット」「社内リンク先」「社外リンク先」などを実装することとした。とくに、「単語セット」は知識の構造化論についての多くの先行研究が重視している「ノード」と「リンク」という概念を取り入れて、そのままでは意味発見が難しいオリジナルデータから、意味発見が容易になるように工夫することとした。

第二の統計解析手法としては、出現頻度の少ない潜在的な重要単語を発見するための手法の確立を目指した。そのために、上述の「単語セット」を起点として、それを「発散」と「収束」を繰り返して進化させる手法を提示し、さらにこのようにして進化した単語セットの中からコミュニティを抽出したり、ネットワーク類似性を発見したりすることによって、時系列で見た技術開発の趨勢の発見のための手法や、潜在的な有望技術・ニーズを発見する手法を提示することを目的とした。

以上の目的を達成するうえで、この二つの手法が有効であることを実証するために、本研究ではいくつかの産業・技術分野をとりあげ、その分野の特許公報、学術論文、ウェブ上で公開されたプレスリース等を収集した。収集したのは、(1)電気自動車、太陽電他、サービスロボット等にかんする日本特許庁の電子図書館からダウンロードされた特許公報、(2)科学技術振興機構が提供している学術論文のデータベースである JDreamⅢに収録された論文の抄録、(3)『日本ロボット学会誌』『ＰＭ学会誌』などのフルテキストをスキャナーで読み取り、テキストデータに変換したフルテキストデータ、(4)インターネットの企業のサイトやニュースサイトに公開された太陽電池、生物模倣(バイオミミクリ)などの分野のウェブデータなどである。

次いで、これらのデータをリレーショナルデータベース、および Content Analytics などテキストマイニングツールにインポート可能であるように、csv ファイルとして加工

する。この加工に当たっては、時系列の情報を得ることが出来るように各データが発表された年次を可能な限り明記するように努め、また著者所属機関・国籍などのデータも抜け落ちないように努めた。

このようにして得られたテキストデータは、そのオリジナルなままで統計解析するのが従来のテキストマイニング手法であったが、本研究ではこのデータにまず付加価値を与えることを重視している。したがって、実際にSalesforce社のリレーショナルデータベースを利用して知識創出・共有システムを試作するとともに、そのようにして作成され付加価値のついたテキストデータをテキストマイニングツールを用いて解析した。本研究の事例研究として選択されたのは、太陽電池、電気自動車などの分野であり、利用したテキストマイニングツールは、「Contents Analytics」および「Text Mining Studio」である。この解析ツールにより、形態素解析をおこない、形態素の出現頻度の時系列データ、および形態素の共起頻度を示す共起行列を得て、このデータに多変量解析やネットワーク分析を適用することによって、このようなアプローチなしには得られなかっただろうはずの新たな知見を発見することを目指した。

４．研究成果

本研究の成果は以下の通りである。

第一の柱であるデジタル空間上でデジタルデータベースにアノテーションの付与し、知識創出・共有システムを構築するための工夫については、リレーショナルデータベースの中に、「単語セット」「アンダーライン」「コメント」「内部リンク先」「外部リンク先」等の属性を作成し、実験的に書く属性に書き込みをおこなった。

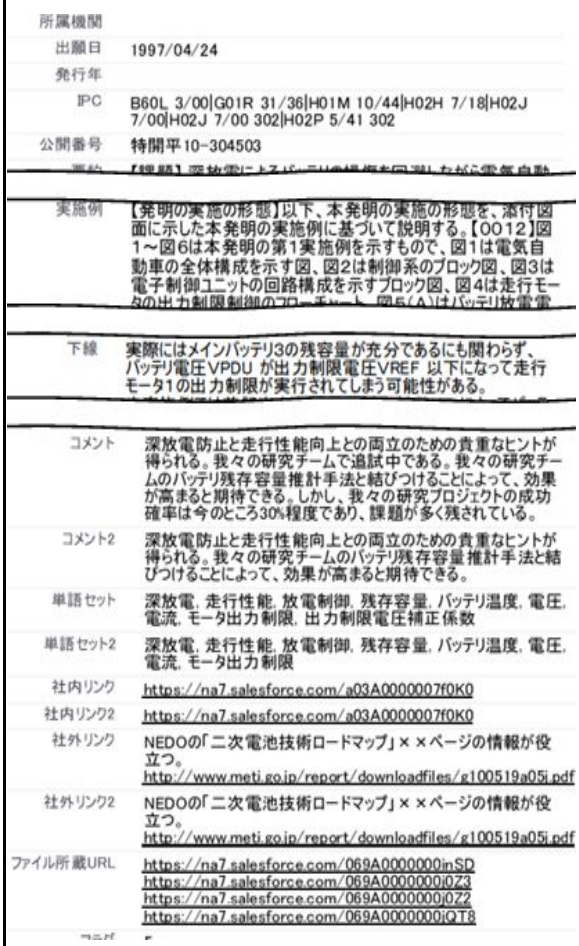
とりわけ重視されるのは単語セットであり、これに注目することによって重要な知見や意味を含んでいるテキストデータを絞り込むことができること(＝無意味なノイズを除去)、および単語セットを不断に進化させることによって、テキストデータから出現頻度や共起行列の精度を高めることができることを実証した。

単語セットには、オリジナルテキストデータの中に含まれている知識がこのシステムのユーザーの脳の中に構築されている知識体系に合う形で作成され、書き込まれる。そしてこのシステムが利用を続けることによって、この単語セットは進化される。また、本システムの他のユーザーもこれから重要なヒントが得られることを考慮して作成される。

さらに、コメント欄にはオリジナルデータに対するユーザーの考え(肯定的および否定的な意見、このデータの深い意味や別の解釈等々)を書き込むことにする。さらに「リンク先」にはこのテキストデータの意味を理解する上で参考となるテキストデータやその他の情報のURL等を自動生成して貼っておく。

このシステムを実際に試作するためには、専門的な知識を持つ実際のユーザーを想定して実施することが必要であったので、専門家の協力を得ておこなった。

その一つの事例は以下の図に示される。



所属機関	
出願日	1997/04/24
発行年	
IPC	B60L 3/00|G01R 31/36|H01M 10/44|H02H 7/18|H02J 7/00|H02J 7/00 302|H02P 5/41 302
公開番号	特開平10-304503
実施例	【発明の実施の形態】以下、本発明の実施の形態を、添付図面に示した本発明の実施例に基づいて説明する。【0012】図1～図6は本発明の第1実施例を示すもので、図1は電気自動車の全体構成を示す図、図2は制御系のブロック図、図3は電子制御ユニットの回路構成を示すブロック図、図4は走行モータの出力制限制御のフローチャート、図5(A)はバッテリ放電電
下線	実際にはメインバッテリ3の残容量が充分であるにも関わらず、バッテリ電圧VPDU が出力制限電圧VREF 以下になって走行モータ1の出力制限が実行されてしまう可能性がある。
コメント	深放電防止と走行性能向上との両立のための貴重なヒントが得られる。我々の研究チームで追試中である。我々の研究チームのバッテリ残存容量推計手法と結びつけることによって、効果が高まると期待できる。しかし、我々の研究プロジェクトの成功確率は今のところ30%程度であり、課題が多く残されている。
コメント2	深放電防止と走行性能向上との両立のための貴重なヒントが得られる。我々の研究チームのバッテリ残存容量推計手法と結びつけることによって、効果が高まると期待できる。
単語セット	深放電, 走行性能, 放電制御, 残存容量, バッテリ温度, 電圧, 電流, モータ出力制限, 出力制限電圧補正係数
単語セット2	深放電, 走行性能, 放電制御, 残存容量, バッテリ温度, 電圧, 電流, モータ出力制限
社内リンク	https://na7.salesforce.com/a03A0000007f0K0
社内リンク2	https://na7.salesforce.com/a03A0000007f0K0
社外リンク	NEDOの「二次電池技術ロードマップ」××ページの情報が役立つ。 http://www.meti.go.jp/report/downloadfiles/g100519a05j.pdf
社外リンク2	NEDOの「二次電池技術ロードマップ」××ページの情報が役立つ。 http://www.meti.go.jp/report/downloadfiles/g100519a05j.pdf
ファイル所蔵URL	https://na7.salesforce.com/069A0000000inSD https://na7.salesforce.com/069A0000000j0Z3 https://na7.salesforce.com/069A0000000j0Z2 https://na7.salesforce.com/069A0000000iQT8

この試作実験から得られた重要な成果の一つは、このようなアノテーションは他者との共有よりも、個人的に利用されるときに有効性が増すことが多いことが分かり、したがってその仕組みは共有を重視するよりもむしろパーソナライズ化を優先することが望ましいのではないかということであった。このような知見は、従来の多くの企業が導入したにもかかわらず、十分に成果をあげることができなかったと言われることが多いいわゆるナレッジマネジメントの実践の経験と符合している。

この成果は、電子情報通信学会での研究報告や、科学技術振興協会（ＪＳＴ）の研究誌に発表され、反響を得ることができた。

本研究の第二の柱は、第一の柱の成果を受けて、付加価値のつけられたテキストデータにテキストマイニングツールを適用し、それによって作成された共起行列を統計解析することである。

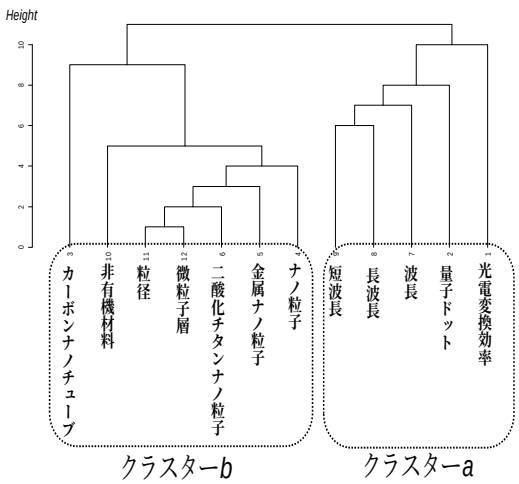
統計解析においては、単語セットに注目し、単語セットを進化させることによって精度を高め、ノイズを除去して核心的な箇所にフォーカスできるような工夫おこなった。具体的には、現時点で得られている単語セットを起点として、それにアソシエーションルールを適用し、単語セットを豊かにすることを目指した。

これによって単語セットを豊かで精緻化した後で、その単語セットに含まれる単語を重視して、単語の出現頻度の時系列データと共起行列を解析した。

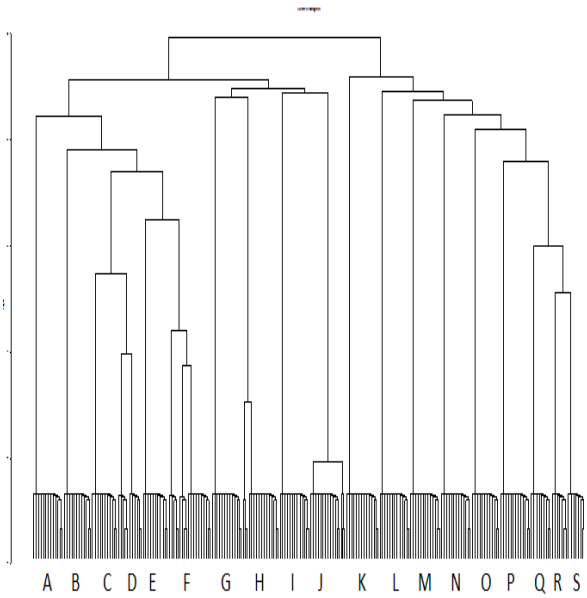
時系列データについては、ネットワーク分析の「ネットワーク類似性」に着目し、単語の共起関係から推定されたネットワーク間の類似性が時間とともにどのように変化するかから、技術進歩の動向を見いだすための工夫を行った。その一事例は以下の図に示されるように、電気自動車の技術進歩が2007-08 年に大きな変化を示したのではないか等の発見をおこなった。

	95-96年	97-98年	99-00年	01-02年	03-04年	05-06年	07-08年
95-96年	1	0.43822	0.43939	0.44752	0.3435	0.45001	0.13299
97-98年	0.43822	1	0.64261	0.57776	0.45841	0.57086	0.22381
99-00年	0.43939	0.64261	1	0.60158	0.4582	0.57794	0.21083
01-02年	0.44752	0.57776	0.60158	1	0.45038	0.59034	0.219
03-04年	0.3435	0.45841	0.4582	0.45038	1	0.53797	0.18325
05-06年	0.45001	0.57086	0.57794	0.59034	0.53797	1	0.24338
07-08年	0.13299	0.22381	0.21083	0.219	0.18325	0.24338	1

また、単語セットを的確に進化させることによって、太陽電池の技術体系が以下の図のように不十分なクラスター分析結果から、次第にコミュニティａが「構造」「形態」「形状」のような単語を取り込んで進化し、太陽電池技術にとって基本的なコンセプトの追求が大切であること等の知見が得られることが発見できた。



さらに、共起関係を二項関係としてとらえるだけでなく、共起関係と共起関係との間の相関を見ることによって、二項関係を超えた単語の共起関係を知ることができるという提案も行った。たとえば、下図のクラスターA には(算出;実現,正電極,電費,充電状態)(電費;算出,温度,電圧)(命令;可能,CPU,センサ,温度)(電圧;電費,正電極)(温度;電費,命令) (劣化;内部抵抗,演算)という関係が含まれているので、これから「「充電状態」を正確に「算出」することによって「電費」の向上を「実現」できた」という意味を類推できるということを示した。



これらの成果は上述の学会報告などのみでなく、二冊の著作(共編著『特許公報のテキストマイニング』(ミネルヴァ書房)、『技術戦略としてのテキストマイニング』(中央経済社))含む多くの研究として発表された。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計３件）

① 菰田文男「「単語セット」の作成と進化に基づくテキストマイニング手法」『情報管理』54 巻 9 号、2011 年、査読有、pp.568-578
② 山本真照・菰田文男「知識共有システムを利用したテキストマイニング手法」『電子情報通信学会技術研究報告』NLC2011-11、2011 年、査読無、pp.55-60
③ 菰田文男「都市とイノベーション」『計画行政』33 巻 4 声、2010 年、招待論文、pp.15-20

〔学会発表〕（計１件）

① 山本真照・菰田文男「知識共有システムを利用したテキストマイニング手法」電

子情報通信学会テキストマイニング部会、2011 年 7 月 8 日、日本アイビーエム本社

〔図書〕（計２件）
① 菰田文男・那須川哲哉編『技術戦略としてのテキストマイニング』中央経済社、2014 年、296 ページ
② 豊田祐貴・菰田文男編『特許情報のテキストマイニング』ミネルヴァ書房、2011 年、274 ページ

〔産業財産権〕
○出願状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
菰田文男（KOMODA, Fumio）
埼玉大学・経済学部・教授
研究者番号：60116720

(2)研究分担者
木戸冬子（KIDO, Fuyuko）
東京大学・情報理工学研究科・助教
研究者番号： 60527828

(3)連携研究者
（　　）

研究者番号：