大規模文書ネットワークからの関連情報抽出
Information retrieval from a large-scale document network

要　旨

オフィスに蓄積された大量の文書を、その属性に基づいて結び付けることによりネットワークとして構造化し、それを分析することにより有用な情報を見出す研究を進めている。分析の方法は、脳の神経細胞（ニューロン）のネットワークが、ある事象から関連する他の事象を連想想起する機構をモデルに考案したアルゴリズムに基づく。このアルゴリズムの基本性能を評価し、他のアルゴリズムに比べて優位であることを示した。さらに、特許間の引用関係が成すネットワークから提案アルゴリズムを用いて関連特許を抽出するシステムを構築した。このシステムを特許調査の実例に適用し、過去に特許調査者によって行われた調査と同等の結果が得られることを確認した。

Abstract

In the average office, a variety of documents are accumulated daily. A large-scale network is gained when these documents are linked by their attributes (for instance, similarities or citation relations). Here, we propose a method to extract relevant information from such document networks. This method is based on the algorithm inferred from the brain mechanism for retrieval of associative memory. We demonstrated the validity of this method by applying the algorithm to a citation network of patents. A set of patents retrieved by this method was in good agreement with those arranged by human experts. These results suggest that the proposed method might be useful for finding valuable information in a pile of office documents.

執筆者

園田　隆志（Takashi Sonoda）
岡本　洋（Hiroshi Okamoto）
坪下　幸寛（Yukihiro Tsuboshita）

研究技術開発本部　システム要素技術研究所
（System Technology Laboratory, Research & Technology Group）

1. はじめに

我々の周りには大量の情報が存在する。机の上には、たくさんの書類が積んであり、脇にはPC が置いてある。書類は会議で配布された資料であり、読みかけの論文である。PC の中には、作成中の原稿や、過去の報告書が格納されており、ネットワークを通して共通サーバーのファイルにもアクセスすることができる。さらに、脇の本棚には書籍や雑誌を置いている。我々は、あらゆる情報を利用することが可能なように思える。しかしながら、これらの情報は、毎日増え続け、整理されず格納されていることも多く、必要な情報を必要なときに利用できないこともしばしばである。過去に作った文書を再利用しようとして、探し始めるが、どこに保存したか分からなくなり、新たに作成することはよくあることである。

ところで、インターネットの世界には大量の情報が Web ページとして存在する。これらの情報は誰かの手によって整理されたものでないにも係わらず、様々な検索サービスにより、過去に閲覧したことのあるページや、新しく作成されたページを見つけ出すことができる。我々が Web ページを利用する場合、リンクを辿って、欲しいページを探してゆく。したがって、リンクされている数が多いページほど、多くの人に見られる可能性が高い。Google™検索エンジンに利用されている PageRank™ アルゴリズムは、このリンク構造を利用して統計的に Web ページの順位付けを行っている[1]。このように、我々の周りの大量情報にも Web ページのリンクのような関係を付けることができれば、あらかじめ整理しておくこと無しに、検索や再利用を行うことができると考えられる。

この Web ページとリンクの関係を抽象化すると、点と線からなるグラフ構造を見出すことができる。このようなグラフはネットワークと呼ばれ、自然の中に数多く存在する。たとえば、人と人の知り合い関係や、鉄道における駅と路線などの関係である。また、脳も多数の神経細胞が結合するネットワークであることが知られている。目や耳から情報が入力されると対応する神経細胞が活性化し、ネットワークを通して、活性を伝播し、関連する情報が想起される。この想起の仕組みが理解できれば、情報のネットワークに適用し、情報の検索ができると考えられる。ここから、我々は、オフィスの文書を、その文書が持つ属性でネットワーク化し、脳の神経細胞ネットワークの情報処理に基づくアルゴリズムによって、有用な情報を得ることを目的に研究を進めている。

オフィスには様々な文書が存在し、それらはまた、様々な属性を持っている。基本技術の獲得のためには、曖昧さのない属性を持ち、評価が可能な文書から出発したい。そこで、我々は、特許を例として研究を進めることとした。特許は、権利化のための文書であり、曖昧さはできる限り排除されている。さらに、我々技術者の身近な文書であり、その評価も技術の重要さとして可能である。特許は国内だけでも毎年 40 万件が特許庁に出願されている。その中から、自身の研究開発と関係のある特許を見出すことは、我々、技術者にとって重要な仕事である。このように特許を対象とすることは、技術的な面だけでなく、効用としても意味があると考えている。

本論文では、大量の特許を審査官引用で結び付け、脳の情報処理からヒントを得たアルゴリズムで分析することで、重要な特許の発見や、技術領域の全体像を知る方法を紹介する。本節に続く第 2 節では、脳の情報処理に基づく大規模ネットワークからの関連情報検索について説明する。第 3 節で、アルゴリズムの紹介と、基本性能の評価を行う。第 4 節では、特許における属性を審査官引用とした文書ネットワークについて説明し、さらに、実例に適用し特許調査における有効性を検証する。

2. 大規模ネットワークからの課題依存的コミュニティ抽出

文書間の関係が構成する大規模ネットワークの全体を個々のユーザに見せても、さほど有益ではなかろう。各々のユーザは固有の課題を抱えており、切に望まれるのは、個々のユーザ課題に応じた情報をネットワークから抽出・生成することであろう。例えば、全体の中からユーザ課題に関

連する文書群をそれらの間の関係性とともに取り出すということである。こうしたことにより、その課題を解決するためにはどのような文書を読めばよいか、さらに、これら文書の間の関係はどうなっているかが把握できる。

我々はこのような要望をかなえるべく、大規模ネットワークからユーザ課題に応じたコミュニティ（ネットワーク科学では意味のある一塊として抽出された部分ネットワークのことを「コミュニティ」と呼ぶ）を抽出し、さらにコミュニティを構成する個々のメンバー（部分ネットワークのノード）の重要度を計算するアルゴリズムを提案した。このアルゴリズムは、脳が記憶を想起する機構をモデルに考案された。

心理学ではしばしば、「記憶」を２つ側面から論じる。「長期記憶」と「短期記憶」である。「長期記憶」とは、個体が誕生以来の様々な経験を通じて獲得した膨大な数の観念が蓄積・保管されたものである。一方、「短期記憶」とは、日々直面する様々な状況に応じて、長期記憶の中の関連する部分が一時的に活性化されたものである。長期記憶を「舞台」にたとえるならば、短期記憶とはそこで上演される劇の時々刻々の「場面」に相当する。

長期記憶は脳内において、観念の間の連想関係が成すネットワークとして構造化されていると考えられている[2]。そうであるならば、短期記憶の想起とは、長期記憶のネットワークから、目下直面する状況を引き金に連想関係を通じて活性化された観念の群が成す部分ネットワーク、すなわち、コミュニティを抽出する過程に他ならない。例えば、火を目撃することにより、「火」に連なる「赤」、「消防車」、「救急車」、「家」、…、といった観念が活性化される。活性化されたこれらの観念が構成するコミュニティは、＜火事＞という概念を表すであろう。

そこで、脳における長期記憶のネットワークを実世界における様々な大規模ネットワーク（身近な例では、WWW/インターネット、ソーシャルネットワークサービス（SNS）における知人関係のネットワーク。その他、文書間の引用関係、生体内の複雑な代謝反応系、等もネットワークを構成する）に対応させてみよう。そうすることにより、脳が状況に応じて長期記憶から短期記憶を読み出すのと同様な機構を通じて、これらの大規模ネットワークから関連情報を引き出せると期待される（図 1）。

脳が長期記憶から短期記憶想起を読み出す機

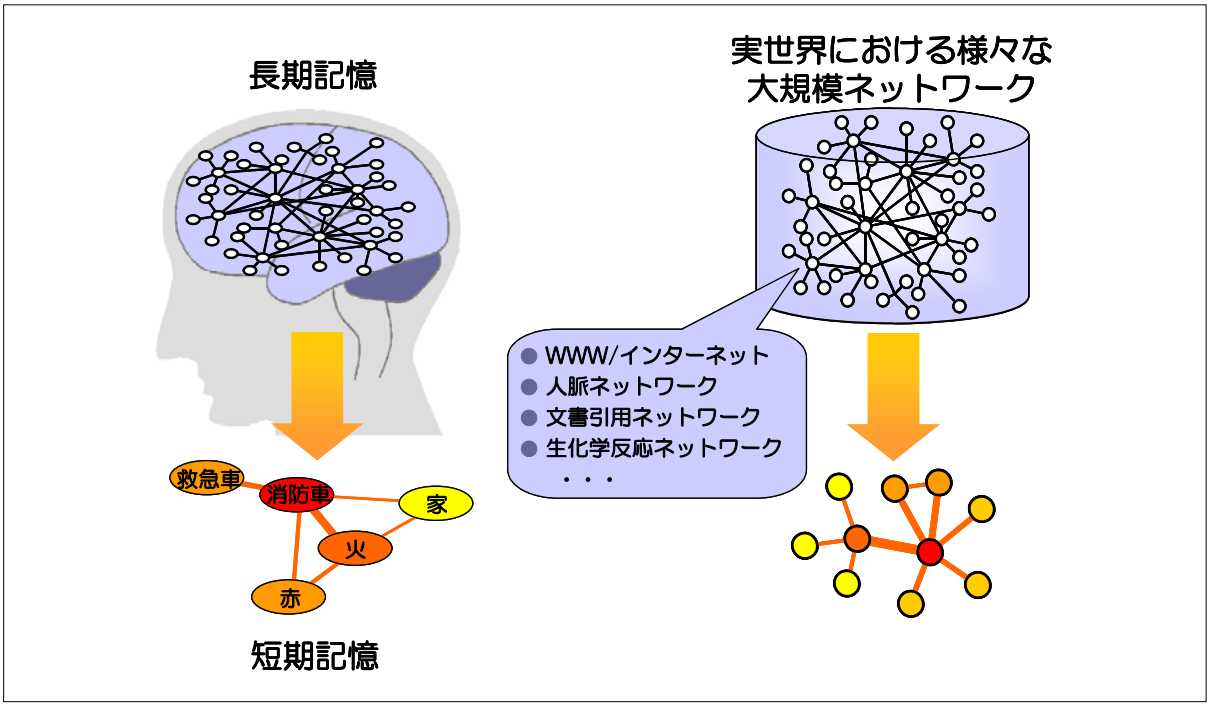

図 1.　脳における長期記憶からの短期記憶読み出しとのアナロジーとしての大規模複雑ネットワークからの関連情報抽出
Information retrieval from large-scale networks in analogy with short-term memory retrieval from long-term memory in the brain.

構の解明は、現在の神経科学における中心テーマの一つであり、我々もこの研究に携わっている。全貌解明にはまだ時間がかかりそうである。しかしながら我々は、「漸次的持続活性」の現象に関する神経生理学的[3-5]・計算論的知見[6-8]に基づき、この機構の（少なくとも一面の）本質が「連続アトラクター力学」で表現できるという仮説を導いた[9-12]。提案アルゴリズム、すなわち、大規模ネットワークから課題依存的にコミュニティを抽出し、さらにコミュニティを構成する個々のメンバーの重要度を計算するという方法は、連続アトラクター力学により記述される。第３節でこのアルゴリズムの詳細を述べるが、数学的記述に興味のない読者は、直接第４節に飛んでもらってかまわない。

3.1　連続アトラクター力学

隣接行列 $\mathbf{A}$ で定められる大規模ネットワークが与えられたとする。$\mathbf{A}$ の ij 成分 $A_{ij}\ (i,\ j=1,\cdots,N)$ はノード j からノード i へのリンクの強さ（重み）である。自己結合はないとする($A_{ii}=0$)。ネットワークの個々のノードにニューロンを、個々のリンクにニューロン間結合（シナプス結合）を対応させる。

各ニューロンは連続値で表される活性を持つ。ニューロン i の活性を p_i とする。ニューロン間でリンクを介した活性の受け渡しが行われる（活性伝播）。ニューロン i への入力 I_i は他ニューロンの活性の線形和で与えられる：

$$I_i=\sum_{j=1}^{N}T_{ij}\,p_j$$ 。ここで、 $$T_{ij}=A_{ij}\Big/\sum_{i=1}^{N}A_{ij}$$

である（T_{ij} は PageRank™ アルゴリズム[1]における遷移確率行列に相当する）。I_i から p_i への変換は、図 2a に示す多重ヒステリシス入出力関係で定められる。多重ヒステリシス入出力特性は、漸次的持続活性を説明するためのニューロンモデルの性質として提案された[13-16]。

１個のニューロンに多数のヒステリシスループが対応するために、その入出力計算には、多数の力学変数が必要となる。そこで、シミュレーション計算の負担を軽減するために、連続極限 $\delta\to 0$ をとる（図 2b）。この極限では、時刻 t における活性パタン $\vec{p}(t)=\left(p_1(t),\cdots,p_N(t)\right)$ から１ステップ先の時刻 $t+1$ における活性パタン $\vec{p}(t+1)$ を、次の簡単な規則で計算できる：

i)　If $I_i(t)<H_1$,　$p_i(t+1)=\alpha_1 I_i(t)$.

ii)　If $H_1\le I_i(t)\le H_2$,　$p_i(t+1)=p_i(t)$.

iii)　If $H_2<I_i(t)$,　$p_i(t+1)=\alpha_2 I_i(t)$.

ここで、α_1、α_2、H_1 および H_2 は、図 2b 中で定義されるパラメータである。

この規則に従ってネットワーク活性伝播が行われると、平衡状態における活性パタン $\vec{p}$ は初期状態（$t=0$ における活性パタン）に連続的に依存する、すなわち、「連続アトラクター」が生成される。この性質により、初期状態として表現された課題に依存した情報読み出しが実現さ

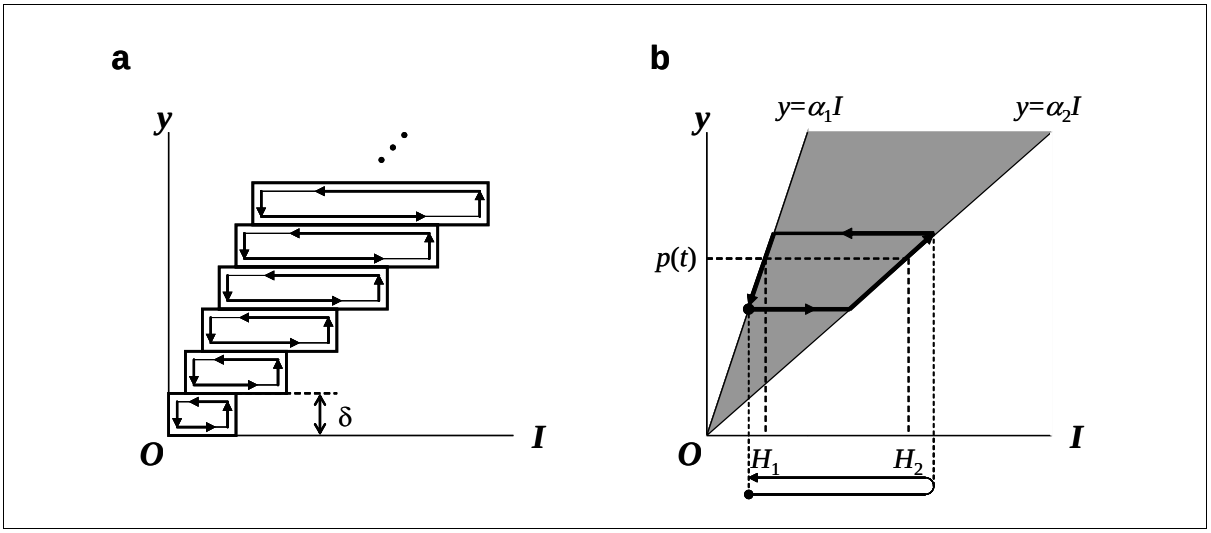

図 2.　a, 多重ヒステリシスニューロンの入出力特性。b, 連続極限（$\delta\to 0$）。
a, Input/output relation of a neuron with multiple hysteresis loops. b, Continuous limit（$\delta\to 0$）.

れる。$\alpha_1 = \alpha_2 = 1$、すなわち、ヒステリシスをゼロとすると、平衡状態は初期状態の情報を消失する。このように、連続アトラクターの生成、すなわち、課題に依存した情報の読み出しには、ステリシス特性が本質的である。ヒステリシス($\alpha_1 - \alpha_2 > 0$)のおかげで、連続アトラクターは微小な外乱に対して安定である。

3.2 アルゴリズム

ユーザ課題を、次のように、ネットワークの初期活性パタンで表す。初期の時点でユーザが知る（不完全な）知識に対応するノードに活性を与える。例えば文書間関係ネットワークの場合であれば、ある課題に対して、ユーザがとりあえず初期の時点で知る関連文書に対応するノードに活性を与える。

活性伝播が導く連続アトラクター状態において活性化されたニューロン群が関連情報を表すと考える。活性伝播を通じて、リンク関係から不要とみなされたノード（活性を失ったノード）は削除される。一方、リンク関係から必要とみなされたノード（高い活性を付与されたノード）は現れる。このような削除・付加の過程を通じて、ユーザが持つ不完全な知識（初期状態）から、ユーザが本来知るべき知識（連続アトラクター）が想起される。

連続アトラクター状態において個々のノードが獲得した活性を、対応するノードの「重要度」と考え、その大きさに従ってノードをランク付する。各ノードへの入力はそれにリンクする他ノードの活性の線形和となっているので、多数からリンクされるノードは概ね高い活性を示す。しかしながら、例えば、同じ数のノードからリンクされる２つのノードでは、より高い活性を持つノードからリンクされる方が高い活性を示す。このように活性値は、被リンク数[17]そのものより、個々のノードの重要度を表す指標として、適切であると考えられる。

ところで、この重要度計算の方法は、一見Google™ 検索エンジンに用いられているPageRank™ アルゴリズム[1]に似ている。しかしながら、決定的な違いがあることに注意したい。個々の Web ページの重要度を確率伝播の平衡状態として計算する PageRank™ アルゴリズムでは、より多くのより重要なページからリンクされているページにより高い重要度が付され、そこにはユーザ個別の興味・課題は反映しない。しかしながら実際には、ある文書はある課題の下では中心的役割を果たすが、別の課題の下では二義的である、という方が普通であろう。我々の方法は、脳が状況依存的に短期記憶を読み出す機構（連続アトラクター力学）をモデルにしているので、課題に依存した重要度の変化を表現できる。なお、$\alpha_1 = \alpha_2 = 1$において、i)-iii)は PageRank™ アルゴリズムに一致する。

3.3 評価：既存アルゴリズムとの性能比較

提案アルゴリズムは「大規模ネットワークからユーザ課題に応じたコミュニティを抽出し、さらにコミュニティを構成する個々のメンバーの重要度を計算する」という機能を実現する。しかしながら、既存アルゴリズムによっても似たようなことができないわけではない。そこで、提案アルゴリズムの優位性を示すべく、既存アルゴリズムとの上記機能に関する性能比較を行った。

定量的な性能比較のための課題を以下のように設定した。Web 検索性能評価のためのテストコレクションとして公開されている TREC 2003 Web Track Distillation task [18, 19]に注目する。これは 50 個の Web 検索課題の集合である。各課題は質問文と正解（Web ページ集合）から成る。それぞれの課題について、正解が「どれだけ繋がっているか」を表す次の量 l [20]を計算する：

$$l = \left[2\sum_{i>j} d_{ij}^{-1} / N(N-1)\right]^{-1} . \qquad (1)$$

ここで、右辺の和は正解を構成するページについてとられる。d_{ij} はノード i とノード j とを最短で繋ぐリンクの数である。i と j を繋ぐ経路がない場合には $d_{ij} = \infty$、すなわち、$d_{ij}^{-1} = 0$ とする。l の大きな課題では、正解 Web ページが繋がっている、すなわち、ネットワークを構成する傾向が強い。このような課題に対しては、Web のネットワーク構造を利用して部分ネットワークを抽出するという方略が有利であると考えられる。そこで、上記機能の評価を、50 個の課題から l がある値 l_θ 以上のものを選んで

作ったサブセットを用いて行うことにした。

提案アルゴリズムにおいて初期活性を付与するノード（種ノード）を、全 Web ページと質問文との語のマッチングにより定める。マッチング上位 M 個を種ノードとした。

線形活性伝播法[21-23]を競合アルゴリズムに設定した。線形活性伝播の考え方は、既存アルゴリズムの中では、提案アルゴリズムの考え方に最も近い。両者はともに、ネットワーク中に活性を伝播させて、その平衡状態として情報を読み出す。両者の違いは課題への依存性をどのように反映させるかという点にある。提案アルゴリズムでは、種ノードに初期活性を与える。一方、線形活性伝播では、活性伝播の過程で種ノードに活性が常にある割合で戻るようにする、すなわち、これらのノードに恒常的にバイアスをかける。線形活性伝播における各ノードの入出力は、その名の通り線形である（図 2b で $\alpha_1=\alpha_2=1$ とした場合に相当する）。しかしながら、種ノードへの恒常的バイアスのおかげで、平衡状態においても課題依存性が維持される。

線形活性伝播は次式で定義される：

$$p_i(t+1)=(1-\beta)\sum_{j=1}^{N}T_{ij}\,p_j(t)+\beta\,E_i/N_S\ . \quad (2)$$

ここで、N_S は種ノードの総数である。ノード i が種ノードであるならば $E_i=1$、そうでないならば $E_i=0$ である。パラメータ β を変化させることにより、どれくらいの割合で種ノードに活性が回帰するかが調節される。線形活性伝播は、ネットワーク上でリンクを辿ってランダムに動くウォーカーが（ここまでは PageRank™ アルゴリズムと同じ）、確率 β で種ノードにジャンプする様子にたとえることができる（PageRank™ アルゴリズムの亜種とみなすこともでき、パーソナライズ PageRank™ アルゴリズムとも呼ばれる[1]）。

PageRank™ アルゴリズムと語の整合との組み合わせによる検索方法[24]との比較も行う。先ず、PageRank™ アルゴリズムを用いて全ての Web ページの「重要度」（ネットワーク確率伝播の平衡状態において、各ノードが獲得した確率で定められる）を求める。先にも述べたように、PageRank™ アルゴリズムにおける重要度はネットワークの構造（隣接行列 $\mathbf{A}$ で定められる）だけから定まり、課題には依存しない。この重要度とは別に、質問文と個々の Web ページとの語の整合度を求める。重要度(p_i)と質問文との整合度(m_i)との係数 γ による重み付き和

$$\mu_i=\gamma\,m_i\Big/\sum_{i=1}^{N}m_i+(1-\gamma)\,p_i\Big/\sum_{i=1}^{N}p_i \quad (3)$$

の大きさに従って Web ページをランク付けし、これを検索結果とする。

3 つのアルゴリズム（提案アルゴリズム、線形活性伝播および PageRank™ アルゴリズムと語の整合との組み合わせによる検索）の性能を次式で定義される”Mean Average Precision (MAP)”で測る：

$$MAP=\sum_{q=1}^{Q_S}AP_q\Big/Q_S\ . \quad (4)$$

ここで、Q_S はサブセットに属する課題の総数、

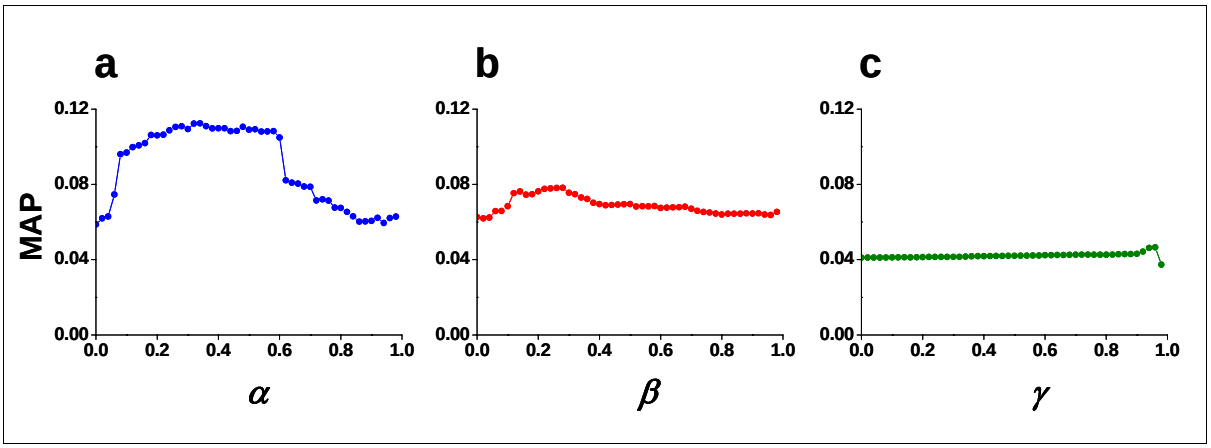

図 3. 3 つのアルゴリズムによる MAP 値。
a, 提案アルゴリズム。b, 線形活性伝播。c, PageRank と語の整合との組み合わせによる検索。
MAP calculated by three algorithms.
a, Proposed algorithm. b, Linear activity propagation. c, Combination of PageRank algorithm and word matching.

AP_q はサブセット中の q 番目課題に関する平均精度であり、次式で与えられる：

$$AP_q = \sum_{i=1}^{n_q} \left(z_i^{(q)} / i \right) \left(1 + \sum_{k=1}^{i-1} z_k^{(q)} \right) \Big/ R_q \,. \quad (5)$$

ここで、R_q は正解ページ総数である。n_q は検索結果としてのページ総数である。検索結果中の順位 i の文書が正解であるならば $z_i^{(q)} = 1$、そうでないならば $z_i^{(q)} = 0$ である。

各アルゴリズムについて、MAP 値を α、β あるいは γ の関数として計算した結果を図 3 に示す。提案アルゴリズムは α の広い範囲（0.1~0.6）において、どの β による線形活性伝播に比べても非常に高い MAP 値を与える。PageRank™ アルゴリズムによる検索は、全ての γ において、提案アルゴリズムおよび線形活性伝播より高い MAP 値を示すことはなかった。以上の結果は、「大規模ネットワークからユーザ課題に応じたコミュニティを抽出し、さらにコミュニティを構成する個々のメンバーの重要度を計算する」という機能を実現する手段としては、提案アルゴリズムが最も優れていることを示す。

4. 特許審査官引用ネットワーク分析

提案アルゴリズムにより実現される上記機能の実際的有用性を、特許の引用関係が構成する大規模ネットワークで示す[25]。提案アルゴリズムを用いることにより、注目する技術領域における発展の系譜あるいはこの領域における各社のパワー関係等、技術に関する調査・分析において、これまで人手で長い時間をかけてようやく得ることができた情報を自動抽出できることがわかる。

ここで、特許における引用について、説明を加えておく。特許にも学術論文と同様に引用文献が付いている。ただし、特許の引用は発明者自身ではなく、特許庁の審査官が付けたものである。進歩性・新規性の視点から、審査官は特許を認めるかどうかの審査を行うが、この過程で審査官が参考にした文献（それらの多くも特許）が特許の引用文献である。発明者には、拒絶理由通知において、進歩性・新規性否定の根拠として、自身が出した特許の引用文献が知らされる。

このように、審査官引用は、引用する側の特許にとって、生死に関わる試練のようなものを表す。従って、特許 A を特許 B が引用することを、A から B への「攻撃」にたとえることができるであろう（図 4 上）。A が B を攻撃した結果、B は撃滅されるかもしれない（拒絶確定）。あるいは、B は防御に成功するかもしれない（答弁書提出後、成立）。

さらに、特許 A の特許 B による引用を、A から B への技術発展の「系譜」とみなすこともできるであろう。上記のように、審査官引用は否定的な意味で付けられることが多い。そこで、引用の系譜を、技術が弁証法的に発展してゆく（先ずは否定され、それが克服されることにより、新しい価値が生まれる）様子ととらえることもできる。けだし、技術の進歩とはそういうものであろう。

審査官引用の連鎖により、特許をノードとし、引用関係をリンクとする巨大なネットワークが構成される（図 4 下）。ところで、企業等の研究・開発において実施される先行技術調査とは、関連する技術群における系譜、すなわち、どれが主あるいは従であるか、どれからどれへの発展が本流あるいは支流であるかを包括的に把握することである。それにより、そこに自身の技術が付加する価値を明らかにできる。また、技術に関する戦略を立案するためには、先ずは、注目する技術領域における競合（この領域で干戈を交える会社）の間のパワー関係の現状、すなわち、「戦況」を知ることが不可欠である。

審査官引用は正にこれらに関わる情報を反映する。従って、特許引用ネットワークから、ユーザ課題に対応する「部分」を、提案アルゴリズムを用いて抽出することにより、この課題に関する先行技術調査あるいは競合間パワー関係の分析に役立つ情報を得ることができると期待される。

ほぼ全ての日本国公開特許（約 1200 万件）の間の審査官引用関係を登録したデータベースを用意する。ユーザ課題を、ユーザがとりあえず知る関連特許群で表現し、これをアルゴリズムに入力する。アルゴリズムはユーザが知るべき関連特許群を抽出する。抽出された特許には、課題に依存した重要度が付される。

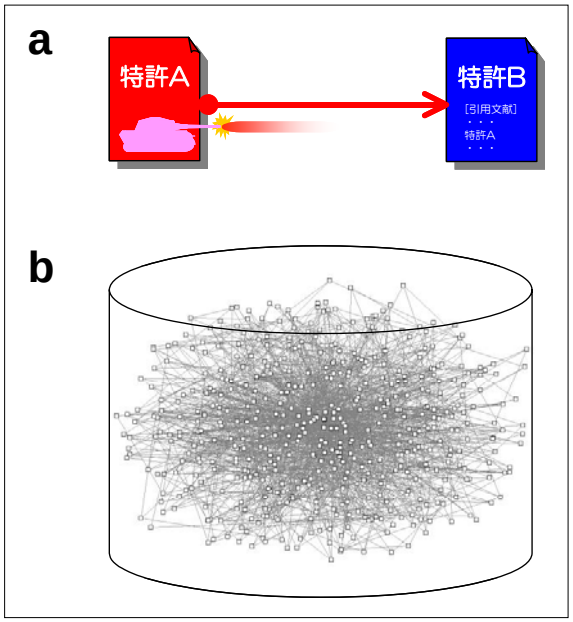

図 4.　特許における引用関係。
a, 特許 A を特許 B が引用。
b, 特許間引用関係が作る巨大ネットワーク。
Citation relation between patents.
a, Patent 'A' is cited by patent 'B'.
b, Large-scale network representing citation relations between patents.

例：「歩行ロボット」に関する関連特許群構造抽出

ここでは、「歩行ロボット」に関する分析結果を紹介する。「歩行ロボット」という語を請求項または要約に含む特許を別途検索し、これらを種特許とした。提案アルゴリズムの処方に従ってコミュニティを抽出し、その構成メンバーである特許を付された重要度の順に並べた。これが歩行ロボットに関する特許の重要度ランキングである。

さらに、抽出されたコミュニティの構造（このコミュニティに属する特許の間の引用関係が構成するのネットワークの構造）を視覚的に把握できるようにするために、これを可視化表示した（図 5）。各特許の重要度を文書アイコンの大きさで表した。出願人（以下では「プレーヤー」と呼ぶ）をアイコンの色で区別した。先に引用関係を「攻撃」に喩えたが、これに対応して、リンクの矢印を引用される側からする側に向け、さらに矢印に攻撃側プレーヤーの色を付けた。これらにより、この技術領域において、どのプレーヤーが重要拠点（重要特許）を制しているか、さらに、どのプレーヤーが攻勢あるいは守勢にあるかが把握できる。技術発展の系譜がわかるようにするために、アイコンを時間順に配置した（上から下に向かって、過去から未来）。

一般にもよく知られている日本の歩行ロボット技術に関する研究・開発の歴史あるいは競合間パワー関係が再現されていることが理解できるであろう（図 5）。例えば、本田技研工業株式会社（ホンダ）が早くから重要特許を数多く保有してこの技術領域における支配力的地位を確立し、現在に至っていることが見てとれる。少し遅れてソニー株式会社がホンダに次ぐ地位を築いている。最近ではトヨタ自動車株式会社が参入を窺っているようである。図 5 のような可

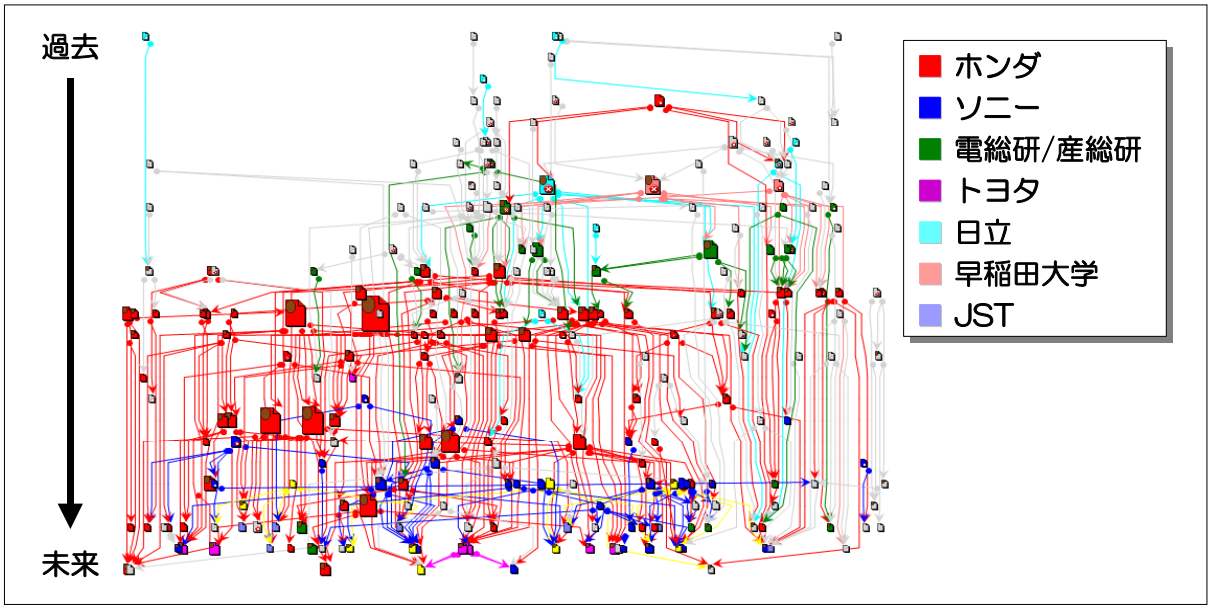

図 5.　「歩行ロボット」に関する提案アルゴリズムによる特許引用分析。
Patent citation analysis by the proposed algorithm for "mobile robot".

視化は、知的財産という無形物を介した戦いを「見える化」[26]するものであり、技術戦略立案のための要図の役割を果たすと期待される。

歩行ロボット以外の様々な課題について、抽出された関連特許群構造の適切さを、個々の技術領域の専門家に評価してもらった。専門家が手作業で調査・分析して把握してきた状況がよく再現できていることを確認した。提案アルゴリズムに基づく特許引用分析に関して、富士ゼロックス株式会社では、単に評価する段階から、開発・研究現場における実際の活動に役立てる段階に入りつつある。

5. おわりに

我々は、大量の文書を、その属性に基づいて結び付けネットワーク化として構造化し、それを分析することにより有用な文書を抽出することを目指している。本論文では、特許の重要さを、技術の先例があるかの判断である審査官引用を利用し分析してきた。さらに、重要さだけではなく、他のことも知りたいとき、何を知りたいかによって結び付けるべき属性も異なってくる。特許調査では、重要な特許を発見するためには、技術者自身が１件１件確認することも必要であるが、F タームで検索しても莫大な数となることが多い。この場合には、関連する特許をグループ化し調査すべき候補を提示できれば、有効だと思える。この目的のためには、属性を引用とするよりは、特許の類似性を表す指標とすることが必要である。本論文では紹介しなかったが、我々の研究チームでは、特許明細書に含まれるキーワードの共通性を近さの指標として特許を結び付け、階層クラスタリング法によって大量特許の分析を行っている。さらに、この分析結果のクラスタ配置をより自然に行うことができる表示技術も開発している[27]。これらの分析方法は、実際に、技術者の特許調査において利用し評価を進めている。

我々技術者にとって特許は重要な情報であるが、同じように論文も重要である。どちらも、新しい発見や発明の報告ではあるが、特許は権利を確保するための報告であり、論文は新しい結果を広く知らせるための報告である。報告の目的が異なれば、そこから得ることができる情報も異なることが推測される。それぞれのネットワークから異なる情報を得られ、組み合わせることでさらに、新しい情報が得られる可能性がある。将来的には複数の種類の対象や、複数の属性のリンクも分析対象としたいと考えている。これによってオフィスに存在する大量の文書の利用が可能になると考えている。

さらに、我々が考える情報は、文書として記述されているものに限らないと考えている。たとえば、文書の作成者や作成した時刻、修正した時刻、さらに、閲覧した者の名前なども属性とすることが出来る。これらの属性は、文書の閲覧や、再利用などによって変化する。そして、これらの変化する属性から、将来の予測が出来ないだろうかと考えている。新しい技術が発展していくときのネットワークの構造変化が法則化できれば、新しい技術の芽の発見が出来ることも可能だろう。このために必要な技術は、時間パラメータを含む大規模ネットワークの分析手法である。

ところで、ネットワーク分析の興味深い点は、個々の要素の振る舞いを定義し、その間の関係を調べると、全体の性質が理解できるということである。個々の振る舞いから全体の性質を見出そうとする試みは、物理学では、統計力学として研究されている。統計力学では、分子のミクロな振る舞いと、システム全体のマクロな性質を結び付ける理論体系である。対象とする系に含まれる要素の数はアボガドロ数程度(10^{23}個)というような莫大な数であるが、近年は、より要素の数が少ない中間的な系も対象とすることができる手法も開発され、インターネットのようなネットワークもその対象となっている[28]。さらに、これまで物理学の対象ではなかった生物システム[29]、機械学習や符号・暗号などの情報理論[30-31]、金融・交通流などの社会システム[32]などの理解にも適用され始めている。人と人の関係を、友人や上司部下などといった属性で結び付けネットワーク化し、組織の性質を明らかにしようとする研究に、社会ネットワーク分析がある[32]。従来は、データの収集はアンケートのような手作業による方法が主であったが、現在は、電子データの利用が可能に

なっており、統計力学的な考えは有効であると考えている。本論分で分析手法の基礎にしている脳も多数の神経細胞から成り立っており、統計力学の対象として研究が進められている[33]。我々もアルゴリズムの評価に統計力学の解析手法を利用し、新しい知見を得始めている。

今後は、文書のネットワークにおいても、統計力学に基づき時間変化を記述する手法を開発し、将来予測も可能にしたいと考えている。現在は、技術者の支援を中心とした課題に取り組んでいるが、さらに、対象文書を拡大し、組織に存在する大量の文書を結び付け、その中から変化する組織に対応した情報を引き出し、オフィスで働く人々のための技術となるように研究を進めて行きたい。

商標について

参考文献

1. Page, L., Brin, S., Motwani, R. & Winograd, T (1998). The PageRank citation ranking: Bringing order to the Web. Stanford Digital Library Technologies Project. http://www-db.stanford.edu/~backrub/pageranksub.ps
2. Collins, A. M. & Loftus, E. F. (1975). A spreading activation theory of semantic processing. Psychological Review 82, 407-428.
3. Romo, R., Brody, C. D., Hernandez, A. & Lemus, L. (1999). Neuronal correlates of parametric working memory in the prefrontal cortex. Nature 399, 470-473.
4. Aksay, E., Gamkrelidze, G., Seung, H. S., Baker, R. & Tank, D. W. (2001). In vivo intracellular recording and perturbation of persistent activity in a neural integrator. Nature Neuroscience 4, 184-193.
5. Egorov, A.V., Hamam, B.N., Fransén, E., Hasselmo, M.E. & Alonso, A.A. (2002) Graded persistent activity in entorhinal cortex neurons. Nature 420, 173-178.
6. Seung, H. S., Lee, D. D., Reis, B. Y. & Tank, D. W. (2000). Stability of the memory of eye position in a recruitment network of conductance-based model neurons. Neuron 26, 259-271.
7. Koulakov, A.A., Raghavachari, S., Kepecs, A. & Lisman, J.E. (2002) Model for a robust neural integrator. Nature Neuroscience 5, 775-782.
8. Miller, P., Brody, C. D., Romo, R. & Wang, X.-J. (2003). A recurrent network model of somatosensory parametric working memory in the prefrontal cortex. Cerebral Cortex 13, 1208-1218.
9. 岡本洋，坪下幸寛，深井朋樹 (2005). 漸次的持続活性の神経生理学的および計算論的知見が開く連想記憶の新しい地平. 日本神経回路学会誌 12, 235-248.
10. Tsuboshita, Y. & Okamoto, H (2005). Information Retrieval Based on a Neural-Network System with Multi-stable Neurons. Lecture Note in Computer Science 3697, 865-872.
11. Tsuboshita, Y. & Okamoto, H. (2007). Context-dependent retrieval of information by neural-network dynamics with continuous attractors. Neural Networks 20, 705-713.
12. Okamoto, H., Isomura, Y., Takada, M. & Fukai, T. (2007). Temporal integration by stochastic recurrent network dynamics with bimodal neurons. Journal of Neurophysiology 97, 3859-3867.
13. Goldman, M. S., Levine, J. H., Major, G., Tank, D. W. & Seung, H. S. (2003). Robust persistent neural activity in a model integrator with multiple hysteretic dendrites per neuron. Cerebral Cortex 13, 1185-1195.

14. Loewentein, Y. & Sompolinsky, H. (2003). Temporal integration by calcium dynamics in a model neuron. Nature Neuroscience 6, 961-967.
15. Teramae, J. & Fukai, T. (2005). A cellular mechanism for graded persistent activity in a model neuron and its implications in working memory. Journal of Computational Neuroscience 18, 105-121.
16. Fransen, E., Tahvildari, B., Egorov, A., V., Hasselmo, M., E. & Alonso, A., A. (2006). Mechanism of graded persistent cellular activity of entorhinal cortex layer V neurons. Neuron 49, 735–746.
17. Garfield, E (1955). Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. Science 122, 108-111.
18. Craswell, N., Hawling, D. (2003). “Overview of the TREC 2003 Web Track”, in the 12th TREC.
19. http://trec.nist.gov/
20. Newman, M. E. J. (2003). The Structure and Function of Complex Networks. SIAM Review 45, 167-256.
21. Shakery, A., Zhai, C. X. (2003). “Relevance Propagation for Topic Distillation UIUC TREC 2003 Web Track Experiments”, in the 12th TREC.
22. Song, R., Wen, J. R., Shi, S. M., Xin, G. M., Liu, T. Y., Qin, T., Zheng, X., Zhang, J. Y., Xue, G. R., and Ma, W. Y. (2004). “Microsoft Research Asia at Web Track and Terabyte Track of TREC 2004”, in the 13th TREC.
23. Tao Qin, Tie-Yan Liu, Xu-Dong Zhang, Zheng Chen, Wei-Ying Ma (2005). “A Study of Relevance Propagation for Web Search”, The 28th Annual International ACM SIGIR Conference (SIGIR'2005).
24. L. Page and S. Brin (1998), The anatomy of a search engine, in: Proc. of the 7th International WWW Conference (WWW 98), Brisbane, Australia, April 14-18.
25. 岡本洋，坪下幸寛 (2008). 特許引用ネットワーク分析：企業競争力源泉としての知的財産権の強化に向けて．情報処理 49, 74-75.
26. 遠藤功　見える化-強い企業をつくる「見える」仕組み. (東洋経済新報, 2005)
27. 武田隼一，池田仁　(2007). 大量特許分類結果の可視化：クラスタの位置および密度分布を考慮した二次元マッピング手法. 第4 回情報プロフェッショナルシンポジウム（INFOPR2007）予稿集, p.67.
28. Dorogovtsev, S. N. & Mendes, J. F. F.(2003). “Evolution of Networks”, Oxford University Press.
29. Sonoda, T. (2002). Adaptation of learning antigens by gene recombination in the immune system, J. Phys. A: Math. Gen. 35, 5973-5983.
30. Engel, A, & Van den Broeck, C. (2001) “Statistical Mechanics of Learning“, Cambridge University Press.
31. 西森 秀稔(1999). スピングラス理論と情報統計力学 (新物理学選書), 岩波書店
32. Volt, J. (2001). “The Statistical Mechanics of Financial Markets”, Springer-verlag.
33. Wasserman, S. &　Faust, K. (1994). “Social Network Analysis”, Cambridge University Press,
34. Amit, D. J. (1989), “Modeling Brain Function“, Cambridge University Press,

筆者紹介

岡本　洋
システム要素技術研究所に所属。
専門分野：理論物理学、計算論的神経科学　(理学博士)

坪下　幸寛
システム要素技術研究所に所属。
専門分野：理論物理学、システム工学

園田　隆志
システム要素技術研究所に所属。
専門分野：理論物理学、機械学習　(理学博士)