活性化拡散モデルに基づく強化学習エージェントの転移学習手法

Transfer Learning Method based on Spreading Activation Model for Reinforcement Learning

○学　高桑優作 (東京電機大)　　河野仁（工芸大）
温文（東大）　　正　神村明哉（産総研）
富田康治（産総研）　　正　鈴木剛（東京電機大）

Yusaku TAKAKUWA, Tokyo Denki University, y.takakuwa@nrl.c.dendai.ac.jp
Hitoshi KONO, Tokyo Polytechnic University
Wen WEN, University of Tokyo
Akiya KAMIMURA, National Institute of Advanced Industrial Science and Technology(AIST)
Kohji TOMITA, National Institute of Advanced Industrial Science and Technology(AIST)
Tsuyoshi SUZUKI, Tokyo Denki University

This paper proposes a policy transfer method of a reinforcement learning agent for suitable learning in unknown or dynamic environments based on a spreading activation model in the cognitive psychology.The reinforcement learning agent saves policies learned in various environments and learns flexibly by partially using suitable policy according to the environment. In the proposed method, an undirected graph is created between policies, and the network is constructed by them. The agent updates the activate value that policy has according to the environment while repeating processes of recall, activation, spreading, attenuation and learns based on the network. Agent uses this network in transfer learning. Experimental simulations comparing the proposed method with several existing methods are conducted to confirm the usefulness of the proposed method. Simulation results show that the reinforcement learning agent achieves task by selecting the optimal one from policies with the proposed method.

Key Words: Reinforcement learning,Transfer learning , Mobile robot, Cognitive psychology

1　緒言

近年，学習能力，認識・理解能力などの人や動物が持つ知的能力を有するロボットの実用化が期待されている．ロボットを知能化し，未知の環境及び動的な環境へ適応を可能にする研究が盛んに行われている [1]．未知の環境や動的な環境へロボットを適応的に行動させるために，人が制御則を与えることは困難であることから，ロボットに自律的に行動を学習させる強化学習，転移学習が用いられている．強化学習とは，エージェント (以下，学習可能なロボットをエージェントと呼称) に試行錯誤を行わせることで，最適な行動を学習させる手法である [2]．また，転移学習は，獲得済み知識を再利用することで，強化学習における新たな環境への適応能力向上や学習時間短縮を図る手法である [3]．強化学習は環境や目的に対して知識 (以下，方策) を獲得可能であるが，エージェントを未知の環境や動的な環境に適応させるために，単一の方策ではなく複数の方策を転移学習により利用させることが必要であると考えられる．

既存研究においても，複数の方策を保存し，選択して再利用した場合の検討がなされているが [4][5]，動的環境や未知の環境における学習は考慮されていないことから，再利用する方策によっては学習効率の低下も予測される．既存研究に対して文献 [7] では，人間の記憶や知識の思い出しや再認識を行うメカニズムである活性化拡散モデルを用いた方策間関係の記述及び方策選択の手法が提案され，計算機シミュレーションにより学習効率の検証がなされた．

本研究では，文献 [7] で用いられた手法を継承し，新たに方策の関連性に基づいたカテゴリを用いて，活性化拡散モデルを再現した転移学習手法を提案する．

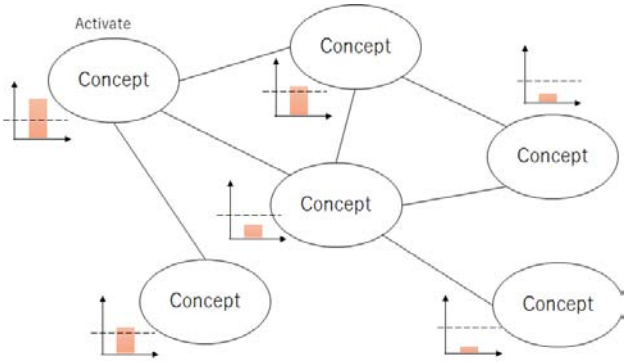

Fig.1 Example of spreading activation model

2　活性化拡散モデル

活性化拡散モデルとは，人間が獲得した概念同士が脳内でネットワーク構造として保存されていることを前提とし，ある概念が想起 (思い出し，再認識等) されることで，関連する概念も活性化され，概念の利用が促進されるモデルである [8]．活性化拡散モデルには，関連性の強さに応じて関連している概念間の距離を変動させて配置する意味的距離が存在する．概念の活性化は，関連性によって構築されたネットワークを通じて行われ，各概念間に意味的な関連性の表現が存在する．活性化拡散モデルの例を図 1 に示す．図 1 では，活性化された概念から伸びる距離を経由して活性値と呼ばれる値が拡散，伝播している様子を表している．本研究では，方策間の関連性に基づきネットワークを構築し，そのネットワークを用いて転移学習を行う．

3 提案手法

3.1 前提条件

本研究では強化学習するエージェントが獲得した方策を選択しながら転移学習を行う手法を提案する．本節では，提案手法を述べるにあたり，関係用語や前提を述べる．強化学習には，Q 学習を用いる．時刻 t での状態 s における行動 a の価値である行動価値 $Q(s_t, a_t)$(以下，Q 値と呼称) の更新式を次に示す．

$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha\{r_{t+1} + \gamma \max Q(s_{t+1}, a)\} \quad (1)$$

s_t, a_t は現在の状態及び行動，s_t は行動 a_t を行った後の状態，α を学習率，γ を割引率，r_{t+1} を報酬とする．

初期位置から目的地までの最短の経路を学習する最短経路問題を Q 学習のタスクとして設定する．学習した Q 値及び学習した環境における環境情報（障害物などの情報）は，Q-table と呼ばれる Look-up table と共に記述しておき，予め再利用可能な状態で保存する．方策再利用に関する転移学習の更新式を次に示す．

$$Q_c(s_t, a_t) \leftarrow \tau Q_s^{\pi_i}(s_t, a_t) + Q_t(s_t, a_t) \quad (2)$$

転移学習では，学習予定のタスク (以下，Target-task と呼称) に対して，予め学習したタスク (以下, Source-task と呼称) で獲得した方策を転移する．$Q_s^{\pi_i}(s_t, a_t)$ は Source-task 方策 (以下，i は方策の識別番号)，$Q_t(s_t, a_t)$ は Target-task で学習中の方策，$Q_c(s_t, a_t)$ は統合した方策を示している．τ は利用方策の Q 値を調整するパラメータ (以下，転移率と呼称) である．

3.2 提案手法の流れ

エージェントは，タスクの達成のため自身の周囲の環境情報を観測し，複数方策を用いて構築したネットワークと観測した環境情報を基に，方策を選択して転移する．この転移学習の目的は，$Q_s(s_t, a_t)^{\pi_i}$ を手掛かりに $Q_t(s_t, a_t)$ を獲得することである．提案手法では，複数方策を用いてネットワークを構築し，想起，方策の選択，活性化拡散，減衰という処理をエージェントの行動毎に反復して行い，方策の活性値を調整しながら転移方策を選択する．

3.2.1 方策ネットワークの構築と方策のカテゴライズ

本研究では、複数方策をカテゴリに分類し，それを基にネットワークを構築する．本研究のカテゴリとは，複数の方策に関連性を見出し，関連性のある方策同士を集めたものを指す．方策同士をある観点で比較し，同じカテゴリに属するかを判定する．この観点のことをプロトタイプと呼称する．このプロトタイプは，予め学習した方策に含まれるデータで利用可能なものを用いる．また，分類したカテゴリ内で，方策間の関連性を記述するための方策間距離 d_{ij} を生成する．方策間距離は，カテゴリ内ですべての方策接続パターンを網羅するように方策同士を全結合する．活性化拡散に基づく方策ネットワーク (以下，SAP-Net : spreading activation policy network と呼称) は無向グラフとし，結合する方策の集合 Π，方策間の接続関係を示す E，距離の持つ重み ω を用いて次式のように定義する．

$$\mathbb{G} = (\Pi, E, \omega) \quad (3)$$

無向グラフの頂点となる方策の集合の元は，$\pi_i \in \Pi$，$e \in E$ であり，E は最大で $E \subseteq \Pi \times \Pi$ である．グラフの表現方法には，Tutte 行列を用いて表現する．Tutte 行列は $n \times n$ の正方行列 M で表され，次式のように行列内の要素を定義する．

$$M_{ij} = \begin{cases} 0 & (\{\pi_i, \pi_j\} \notin E) \\ \omega_{ij} = 1 & (\{\pi_i, \pi_j\} = e \in E) \end{cases} \quad (4)$$

図 2 に SAP-Net の例を示す．図 2 を Tutte 行列で表現すると，次式のように表せる．

$$M = \begin{pmatrix} 0 & \omega_{12} & \omega_{13} & 0 \\ \omega_{12} & 0 & \omega_{23} & \omega_{24} \\ \omega_{13} & \omega_{23} & 0 & \omega_{34} \\ 0 & \omega_{24} & \omega_{34} & 0 \end{pmatrix} \quad (5)$$

SAP-net を構築する際に，分類したカテゴリから生成したプロトタイプ行列によって，方策間距離の重みに影響させる．次式を用いて SAP-Net の各要素の重みを調整する．プロトタイプ行列も SAP-Net と同様に Tutte 行列で表され，式 (4) のようにカテゴリ内の方策間距離の接続から重みを生成する．

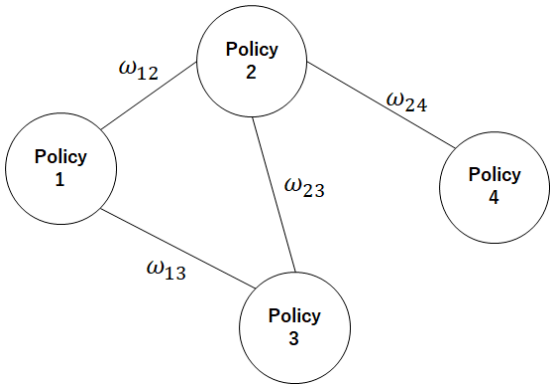

Fig.2 Example of SAP-Net graph

$$M = M + \delta\{\mathbb{P}_1 + ... + \mathbb{P}_n\} \quad (6)$$

n は距離の重複したカテゴリの数を表し，$\delta(-1.0 < \delta < 0)$ はプロトタイプ行列の重みを調節する係数である．

3.2.2 方策の想起

想起では，観測した情報に基づいて，あるカテゴリを選択し，そのカテゴリに分類される方策の中から選択候補を求める処理が行われる．それぞれの方策には，活性値 $\mathbb{A}_i$ というパラメータを与えておき，エージェントは観測情報に基づいてカテゴリを選択することで，カテゴリ内の全方策の活性値を更新する．更新式を次に示す．

$$\mathbb{A}_i \leftarrow \mathbb{A}_i + A_{recall} \quad (7)$$

A_{recall} は，想起係数と呼び活性値の上昇を調節するための係数である．加えて，候補を求める際には，活性値を参照して候補に選択する方策の足切りを行う閾値関数 $\mathbb{T}(\mathbb{A}_i)$ を定義する．次式に示す．H は任意の閾値を示す．

$$\mathbb{T}(\mathbb{A}_i) = \begin{cases} 0 & (\mathbb{A}_i < H) \\ 1 & (\mathbb{A}_i \geq H) \end{cases} \quad (8)$$

3.2.3 活性化拡散モデルに基づく方策選択手法

方策の選択には，エージェントが観測した情報と構築した SAP-Net を用いる．観測情報を用いた想起で得られた方策の選択候補から確率的に方策を選択する．活性値の大きさで選択確率 P_i を算出する．算出式は次式に示す．j は候補方策の識別番号を示している．

$$P_i = \frac{\mathbb{A}_i}{\sum \mathbb{A}_j} \quad (9)$$

3.2.4 活性値の減衰

エージェントが行動する毎に，全方策の活性値を減少させていく．残った活性値を Φ_i とし，次に更新式を示す．λ は減衰を調整する係数である．

$$\Phi_i = \mathbb{A}_i e^{-\lambda} \quad (10)$$

3.2.5 活性化拡散

方策を選択してエージェントが行動した際には，そのフィードバックを活性値に与える活性化という処理を行う．統合方策に従った行動が Target-task の学習を助けるものになった場合 (以下，正の転移) は使用した方策の活性値を増加させる．反対に，学習を妨げる行動につながった場合 (以下，負の転移) は，活性値を減少させる．活性化は，残活性値に活性化係数 $A_{activate}$ を用いて次のように表される．

$$\mathbb{A}_i = \begin{cases} \Phi_i + A_{activate} & (\text{正の転移}) \\ \Phi_i - A_{activate} & (\text{負の転移}) \end{cases} \quad (11)$$

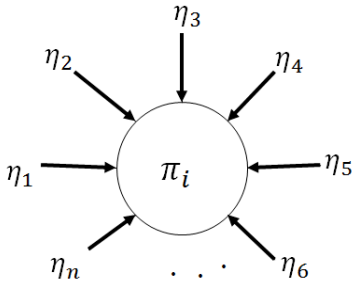

Fig.3 The activation value inputs from neighbor policies via spreading function

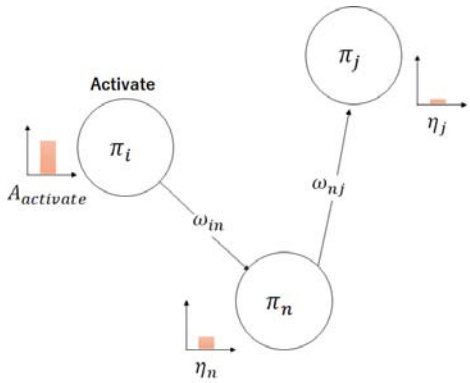

Fig.4 Example of spreaing the activation value inputs

加えて，活性化した方策から伸びる方策間距離を経由して接続関係のある方策へと活性値の拡散処理を行う．拡散について定義するにあたり，SAP-Net 上に保存されている方策 π_i へ拡散される活性値 (以下，活性値入力) の様子を図 3 に示す．ある方策 π_i から活性値入力を，$\eta_k(k = 1, 2, ..., n)$ とし，拡散元 π_i と拡散先 π_j の二つの方策についての活性値の拡散を考えると π_j に拡散される活性値入力 η_j は π_i からの $A_{activate}$ と複数の方策を経由した拡散も考慮し，経由した方策間距離の重みの和 $\sum\omega$，経由した方策間距離の数 h を用いて次式で表す．

$$\eta_j = \begin{cases} 0 & (\sum\omega \geq \omega_{threshold}) \\ \frac{1}{h\sum\omega} A_{activate} & (\sum\omega < \omega_{threshold}) \end{cases} \tag{12}$$

活性値拡散のイメージを図 4 に示す．拡散する範囲の指定をしない限り永続的に拡散可能になるため，経由した方策間距離の重みの和で閾値 $\omega_{threshold}$ を定める．この計算は方策経由毎に再帰的に行い，経由した距離の重みの和が増えることによって拡散される活性値を減少させていく．最終的に，各方策の活性値入力の総和を求め，正の転移と負の転移で場合分けを行い，次式のように活性値を増減させる．

$$\mathbb{A}_j = \begin{cases} \Phi_i + \sum\eta & (\text{正の転移}) \\ \Phi_i - \sum\eta & (\text{負の転移}) \end{cases} \tag{13}$$

4 計算機実験

複数方策を用いた転移学習の効果を検証するため，強化学習による学習，単一方策のみを転移した場合の転移学習，複数方策による提案手法を用いた転移学習を比較する．評価指標は，エージェントのタスク達成に必要となった行動回数 (以下，Step 数) とタスクの達成回数 (以下，Episode 数) で示される学習曲線と全体の学習で必要となった Step 数の総和 (以下，総 Step 数) を用いる．学習曲線と総 Step 数には，各学習結果の 5 回平均した値を用いる．

Table 1 Parameters of experiment

parameters	symbol	value
学習率	α	0.1
報酬	r	1
割引率	γ	0.9
転移率	τ	0.2
プロトタイプ係数	δ	-0.19
初期活性値	$\mathbb{A}_i$	0.1
想起係数	A_{recall}	0.3
想起の閾値	H	0.5
減衰係数	λ	0.09
活性化係数	$A_{activate}$	0.05
重みの閾値	$\omega_{threshold}$	0.75
温度定数	T	0.01

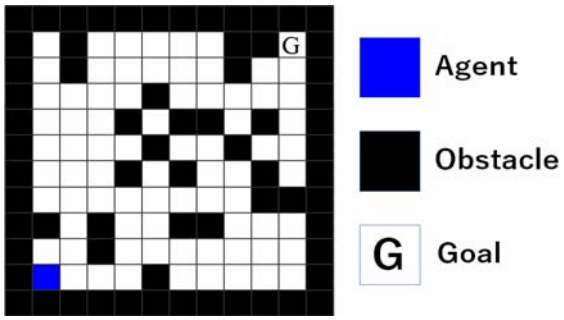

Fig.5 Target-task

4.1 計算機実験の設定

エージェントは上下左右に停止を加えた 5 パターンの行動が可能であるものとし，エージェントの行動選択関数には，ボルツマン分布を用いたソフトマックス手法を用いる．統合方策を参照したボルツマン分布を用いたソフトマックス手法を式 (14) に示す．

$$p(a|s) = \frac{exp(\frac{Q_c(s,a)}{T})}{\sum_{b\in action} exp(\frac{Q_c(s,b)}{T})} \tag{14}$$

実験で用いた強化学習及び転移学習，提案手法のパラメータを表 1 に示す．Episode 数の上限は，1000 回に設定し，使用する方策も同様の回数に設定し学習する．提案手法の正の転移の判定には，エージェントの選択方策の履歴を使用する．目的地へ到達した際には，使用方策の履歴の新しいものに多く活性化がなされるように設定する．次式を用いて使用した方策の活性化を行う．j は方策の識別番号，n は方策使用履歴の順序を降順に並べ替えた時の順序を示している．

$$A_{activate} = \frac{1}{n} \tag{15}$$

エージェントが統合方策に従った行動中に障害物に接触した場合を負の転移と設定する．提案手法で用いる方策数は 100 とする．方策同士が同一カテゴリであるかの判定には，以下のプロトタイプによって判定しカテゴリ毎にプロトタイプ行列を生成する．

- Source-Task 学習開始座標の周囲 1 グリッドの環境情報
- Source-Task の Start 座標から Goal 座標までの方向

エージェントは，行動毎に常に自身の周囲 1 グリッドを環境情報 (障害物，通路) として観測しながら，自身の持つカテゴリと照合する．

学習環境は，グリッドワールドによって構築する．学習に使用する Traget-task の環境を図 5 に示す．予め学習する Source-task の環境は，単一の方策を転移する場合も複数の場合もランダムに構築する．

4.2 実験結果

図 6 に強化学習，単一方策を用いた転移学習，複数方策を用いた提案手法による転移学習の学習曲線比較を示す．RL は強化

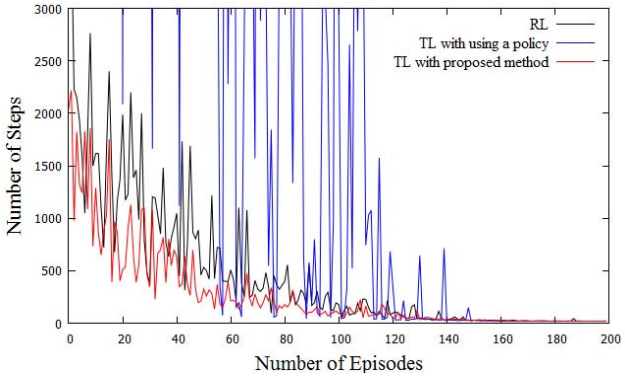

Fig.6 Comparison by learning curve

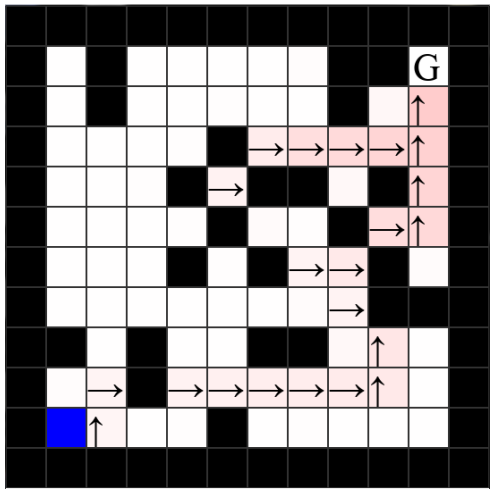

Fig.7 Transfer learning transfered only a single policy

学習を，TL は転移学習を示している．図 6 より，単一方策のみを転移した学習はタスクの達成により多くの Step 数を必要としていることが確認できる．ここで，図 7 に単一方策のみを転移した場合の学習の様子を示す．図 7 の矢印は，その座標で最も行動価値の高い行動方向を指している．この場合，行動価値に従った行動をエージェントが選択すると, 障害物に接触し学習した通りの最短経路を進行することができない．ランダムに学習した方策を転移した場合は，必ずしもその方策が Target-task に適するとは限らないため，図 6 の学習曲線のように方策を使用しない場合よりも学習効率が低下する．この結果が負の転移である．そのため，環境に合う方策の転移の仕方が必要になる．

強化学習と提案手法による転移学習を詳細に比較するため，図 8 に学習 episode 1-100 に注目した学習曲線を示す．図 8 で，学習初期段階における強化学習と提案手法による転移学習を比較すると全体的に提案手法による転移学習の方が Step 数の減少が確認できた．加えて，それぞれの学習における総 Step 数をまとめた表を表 2 に示す．表 2 と図 8 より提案手法を用いた転移学習が総 Step 数が一番少ないことから，学習効率が最も良いことが確認された．ただし，転移学習の特徴である学習初期段階における Step 数の減少 (Jump Start) が観測できなかったことから，Target-task に適する方策が少なかったことが推測される．今回の実験では，ランダムに生成した障害物を配置した環境の最短経路を学習した 100 の方策を使用していることから，Target-Task

Table 2 Total number of steps

学習方法	総 Step 数
RL	110908.4
TL with using a policy	1097578.4
TL with proposed method	72131.2

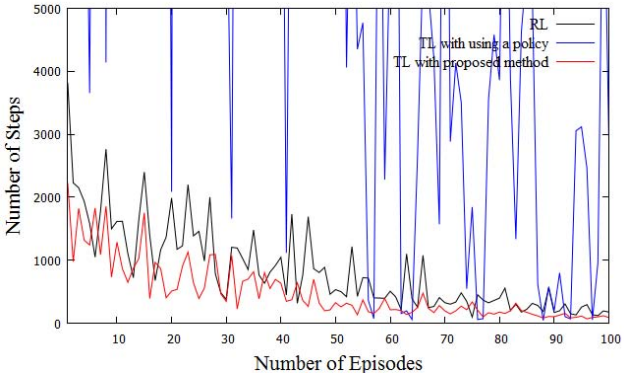

Fig.8 Comparison by learning curve (Episode 1-100)

に適した知識が極めて少ない場合は，今回の結果よりも学習効率が悪くなるケースも考えられる．

5 結言

本稿では，複数方策を用いて構築した SAP-Net とエージェントの観測した情報に応じて選択する手法を提案した．実験では，方策を用いない強化学習と単一方策を用いた転移学習，提案手法を用いた転移学習の学習効率の比較を行い，複数方策を用いた転移学習が有用であることを示した．実験結果で得られた Jump Start の観測が可能な新たな手法や SAP-Net の学習中の最適化などについて検討する．また，提案手法を用いた実機実験や獲得した方策数の学習に対する影響の検証等を行う．

謝辞

本研究の一部は，JSPS 科研費 (16K12493) の助成を受けて行われました．ここに謝意を表します．

参考文献

[1] 赤井直紀，山内健司，井上一道, 宇内隆太郎，山本条太郎, 尾崎功一, “つくばチャレンジ 2013 の課題を材とした実環境におけるタスク遂行ロボットの開発”，計測自動制御学会論文誌,vol.51,No.1,pp24–31,2015.

[2] R.S.Sutton and A.G.Barto，“強化学習”，森北出版株式会社,2000.

[3] M.E.Taylor,”Transfer in Reinforcement Learning Domain”,Springer,2009.

[4] F.Fernandez,M.Veloso,“Probabilistic Policy Reuse in a Reinforcement Learning”,Proceedings of the Fifth International Joint Conference on Autonomous Agents and multi-Agent Systems AAMAS’06,May 8-12,Hakodate,Hokkaido,Japan,2006.

[5] 吉田慎二, 長谷川修,” 強化学習における政策再利用転移学習”, 情報処理学会第 73 回全国大会,2011.

[6] 安藤清志, 石口彰, 高橋晃, 浜村良久, 藤井輝男, 八木保樹, 山田一之, 渡邊正孝, 重野純,” キーワードコレクション 心理学 改訂版”, 新曜社,2012.

[7] 高桑優作, 河野仁, 温文, 神村明哉, 富田康治, 鈴木剛,” 活性化拡散モデルに基づく強化学習エージェントの方策選択手法”, ロボティクス・メカトロニクス講演会講演概要集 2017(0),2P2-E04,2017.

[8] A M.collins,E.F.Loftus,”A Spreading-Activation Theory of Semantic Processing”,Psychological Review,Vol.82,No.6,pp407–428,1975.