ネット情報を使った意思決定のための知識獲得エージェント

藤本 和則　山本 裕

本稿では、インターネット上の情報から意思決定に必要な知識を自動獲得する機構について、これを複数の知識獲得エージェントの協調により実現する協調獲得モデルを提案する。まず、統計的決定に必要な「代替案の集合、情報の集合、確率分布」の知識をそれぞれ獲得する 3 つの知識獲得エージェントについて述べる。そして、これらのエージェントが、容量の限られた通信路を共有しながら知識を収集／獲得する協調獲得モデルを提案する。

1　はじめに

近年、自然言語処理や知識獲得技術の進歩[6] [10] により、インターネット上の情報から様々な知識をある程度の精度で自動獲得できるようになってきた。こうした状況に合わせて、ユーザの意思決定に必要な知識をネット上の情報から自動構成する研究が始められている[1][5]。我々は、特に、不確実な状況下で合理的な意思決定を行うのに必要な知識について、これをユーザとネット情報から自動獲得する意思決定支援の枠組について研究を進めている[8] [9]。

我々は、これまでに、ユーザから「状況の集合と効用関数」を、ネット情報から「代替案の集合、情報の集合、確率分布」をそれぞれ獲得する知識獲得システムの構成を示した[9]。これら 5 種類の知識の獲得により、各代替案の期待効用 (得られる利益の見込み) を計算でき、その大きな代替案を薦めるという意思決定支援が可能となる。本稿では、これらの知識のうち、インターネットを検索しながら獲得する「代替案の集合、情報の集合、確率分布」に着目し、これらを効率的に収集／獲得する知識獲得機構について述べる。

インターネット上の情報から知識を獲得する状況では、獲得に用いる通信路の容量は限られるので、意思決定の質を上げる効果の大きい知識を優先して獲得する機構の実現が重要となる。本稿では、こうした知識獲得機構の一つとして、「代替案の集合、情報の集合、確率分布」のそれぞれを獲得する知識獲得エージェントの協調に基づく協調獲得モデルを提案する。このモデルに基づいて知識獲得を行うことにより、意思決定に用いる知識ベースをより早く価値の高いものとして構成することが可能となる。まず、2 章で、代替案の集合、情報の集合、確率分布をそれぞれ獲得する知識獲得エージェントについて説明する。3 章では、各知識獲得エージェントが意思決定の質を上げるために、協調しながら知識を収集／獲得する機構について述べる。

2　知識獲得エージェント

図 1 に知識獲得システムの構成を示す。図に示すように、知識獲得システムは、ネット上の情報から「代替案の集合、情報の集合、確率分布」を、ユーザから「状況の集合、効用関数」をそれぞれ収集／獲得する。本章では、これらの知識のうち、インターネット上の情報から獲得する代替案の集合、情報の集合、

A Knowledge Acquisition Agent for Decision Making Using Information on the Internet.

Kazunori FUJIMOTO, (有) フジモト・リサーチパーク, Fujimoto Research Park Co., LTD.

Yutaka YAMAMOTO, 京都大学大学院 情報学研究科, Graduate School of Informatics, Kyoto University.

確率分布に着目する。そして、これらを獲得するそれぞれの要素を知識獲得エージェントとみなし、代替案獲得、情報獲得、確率分布獲得のエージェントについて、具体例をもとに説明する。

2.1　代替案獲得エージェント

代替案の集合は、意思決定の対象となる要素の集合であり、例えば、電化製品を購入する場合では、購入対象になる機種が代替案となる。代替案獲得エージェントは、ユーザから意思決定の対象のカテゴリ (例えば「デジタルカメラ」など) が与えられたとき、(1) カテゴリ名をキーワードとして代替案を集めたサイトを検索し、(2) サイトにアクセスしてページを収集し、(3) 集められたページから代替案の集合を抽出する、という過程で獲得作業を実施する。代替案を集めたサイトとしては、例えば、電化製品の選定では価格調査サイト (e.g., http://www.kakaku.com/)、政治家の選定では、議員要覧サイト (e.g., http://www.seiji-koho.co.jp/) などがある。

2.2　情報獲得エージェント

「情報」の集合は、ある代替案を選択した後の「状況」を予測するためのものである。デジタルカメラの購入の例では、この「情報」は、CCD 画素数などの仕様となる。こうした仕様は、機種の画質の良し悪しという「状況」を予測するのに利用できる。情報獲得エージェントは、代替案獲得エージェントの獲得した代替案について、(1) 代替案の名称をキーワードとして「情報」を集めたサイトを検索し、(2) サイトから「情報」に関するページを収集し、(3) 集められたページから「情報」の集合を抽出する、という過程で獲得作業を実施する。「情報」を集めたサイトとしては、例えば、電化製品の選定ではメーカの提供する製品解説ページ、政治家の選定では各政治家のホームページなどがある。

2.3　確率分布獲得エージェント

確率分布は、「情報」の集合を使って、代替案を選択した後の「状況」を確率的に予測するものである。例えば、「情報」、「状況」として、それぞれ CCD 画素数、画質の良し悪しが与えられたとき、この確率分布は、CCD 画素数から画質の良し悪しを予測する確率である。ネット上に直接確率の値が数値として記述されていることはない。そこで、確率分布獲得エージェントは、確率分布に関する知識 (評価知識と呼ぶ) を収集／獲得し、それらを統合して確率分布を導出する (この導出については [2] を参照されたい)。確率分布の獲得は、情報獲得エージェントの獲得した「情報」と、ユーザから与えられた「状況」について、(1) 情報と状況の名称をキーワードとして評価知識を提供するサイトを検索し、(2) サイトから評価知識に関するページを収集し、(3) 集められたページから評価知識を抽出し、(4) 獲得された評価知識を統合して確率分布を構成する、という過程で獲得作業を実施する。評価知識を提供するサイトとしては、例えば、電化製品の選定では、機種の評価ランキングを提供するサイト、あるいは、各機種の個人的な評価を提供する個人ページなどがある。

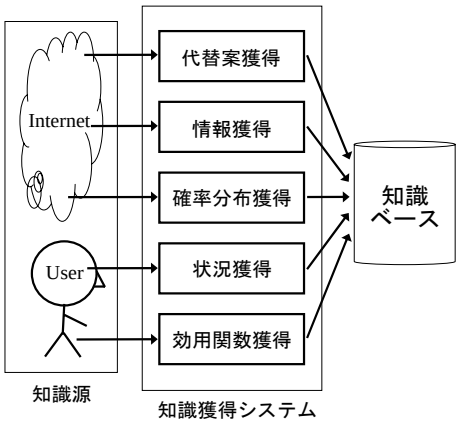

図 1　意思決定系での知識獲得システム

3　協調による知識獲得

本章では、2 章に述べた 3 つの知識獲得エージェントについて、容量の限られた通信路を共有しながら、知識の収集／獲得を行う協調獲得モデルを提案する。

3.1　エージェントの獲得戦略

本節では、知識獲得エージェントの獲得戦略として、知識ベースの価値を増す効果の大きい知識から

優先して収集する戦略を示す。こうした獲得戦略に基づくことにより、知識獲得エージェントは、価値の高い知識ベースをより早く構成することが可能となる。なお、この獲得戦略は各エージェントについて共通となることから、本節では、代替案、情報、確率分布を知識として区別せず、各知識の集合を同一の記号として E、すでに獲得された知識ベースを KB、知識ベースの価値[†1]を $V(\mathrm{KB})$、アクセスの対象となるアドレスの集合を R とそれぞれ表す。また、アドレスへのアクセスには、サイトを検索する行為と、ページを収集する行為が含まれるとする。

以上の設定のもと、エージェントがアドレス $r \in R$ へアクセスしたとき得られる知識ベースの価値の増加量の期待値は、得られる知識 $e \in E$ の見込みを表す確率 $\mathrm{P}(e|r)$ を用いて次式で与えられる。

$$\sum_{e \in E} [V(\mathrm{KB} \cup e) - V(\mathrm{KB})] \cdot \mathrm{P}(e|r) \qquad (1)$$

この期待値をアドレス r の**アクセス価値**と呼ぶ。アクセス価値の大きなアドレスほど、より大きな価値の知識ベースを期待できるという意味で、アクセスする価値があるといえる。こうした考えから、各エージェントは、知識の収集にあたって、得られる価値の期待値のより大きなアドレスを優先してアクセスする。

3.2 協調獲得モデル

協調獲得モデルの構成を図 2 に示す。協調獲得モデルは、知識源、通信路、3 つの知識獲得エージェント、知識ベース、そして、アクセス価値ボードから構成される。アクセス価値ボードは、各エージェントが各知識のアクセス価値を共有するためのメモリである。こうした協調獲得モデルにおいて、各エージェントは、次の 2 つを判断しながら獲得作業を実施する。

- 各エージェント内で、どのアドレスへのアクセスが優先されるか。
- 他エージェントとの間で、どのエージェントが優先されるか。

†1 知識ベースの価値については、様々な定量化が考えられるが、その一例を付録に記す。

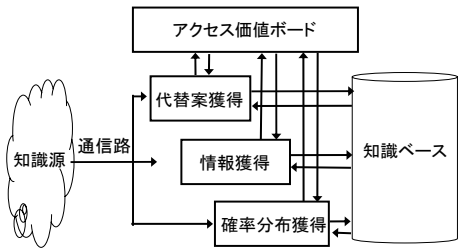

図 2　協調獲得モデル

以下では、通信路の容量制限から、ある時刻に通信路を利用できるのは 1 つのエージェントのみであるとして、効率よく知識獲得作業を実施するエージェントの振る舞いを示す。

1. すでに獲得されている知識ベースを調べ、対象となるアドレスへのアクセス価値を計算
2. アドレス価値の最大値を求め、その最大値を自エージェントの名称と合わせてアクセス価値ボードに提示
3. アクセス価値ボードを調べ、最大の価値を提示するエージェント名をサーチ
4. 最大の価値を提示するのが自エージェントである場合、(1) 収集を実行し、(2) 収集が完了したときアクセス価値ボードから自エージェントのアクセス価値を削除し、(3) 収集された情報から知識獲得を実行し、(4) 知識が獲得された場合は、知識ベースを更新して 1. へ、獲得されない場合は、そのまま 1. へ

こうした振る舞いにより、知識源への通信路の容量に制限があっても、高いアクセスの価値をもつ知識獲得エージェントが優先されながら、効率よく知識ベースを構成することが可能となる。

3.3 関連研究

統計的決定論のもとでの知識の価値については、代替案、確率分布が獲得されていることを前提に、「情報」の価値を議論することが多い [4] [7]。これに対して、図 1 に示す意思決定系では、代替案、情報、確率分布を並行して獲得するので、「代替案の集合、情報の集合、確率分布」の 3 つの要素について、それぞれの価値を考える必要がある。例えば、代替案が全く獲

得されていない場合には、いくら情報や確率分布が獲得されても、利益 (効用) を得ることは期待できない。したがって、知識が全くない場合には、まず、代替案の獲得の価値が高いことになる。一方、代替案のみが獲得され、情報と確率分布が全く獲得されない場合には、適切な選択ができないので、大きな利益を得ることは期待できないことになる。したがって、代替案を獲得した後には、それを選ぶための知識の価値も高まることになる。このように、意思決定系では、「代替案の集合、情報の集合、確率分布」を区別せずに、それらの価値を議論することが重要となる。

情報の価値に基づく獲得戦略については、Value-Driven Information Gathering [3] でもその基本的な考え方が示されている。これに対し、我々の知識獲得モデルは、代替案、情報、確率分布の知識獲得エージェントの協調により知識獲得を実現するという点が大きく異なる。ネット情報から知識を獲得するにあたっては、収集した Web ページ中に知識が存在するかどうかを判定する技術や、表やテキストから知識を抽出する技術など、様々な各知識固有の獲得技術が必要となる。こうした各知識固有の獲得技術をエージェントとしてカプセル化することにより、他知識の獲得技術との独立性を確保でき、技術の追加／変更を容易に行うことが可能となる。また、各知識の獲得手続きの実行に、異なる計算機資源を用いることも容易となる。このように、我々は、提案の協調獲得モデルには多くの利点があると考える。

4　おわりに

本稿では、インターネット上の情報から意思決定に必要な知識を獲得する機構について、これを複数の知識獲得エージェントの協調により実現する協調獲得モデルを提案した。今後は、アクセス価値の定義で用いた「アクセスの結果得られる知識の見込み」の設定法をはじめ、モデルの詳細化を進める。また、今回は、ある時刻に通信路を利用できるのは 1 つのエージェントのみであるとしたが、今後は、容量の限界まで複数のエージェントが利用できるアルゴリズムへの拡張を検討する。

参考文献

[1] Mark Craven, Dan DiPasquo, Dayne Freitag, Andrew McCallum, Tom Mitchell, Kamal Nigam, and Seán Slattery. Learning to extract symbolic knowledge from world wide web. In *AAAI–98*, pp. 509–516, 1998.

[2] Kazunori Fujimoto, Kazumitsu Matsuzawa, and Hideto Kazawa. An elicitation principle of subjective probabilities from statements on the internet. In *Proceedings of the Third International Conference on Knowledge-Based Intelligent Information Engineering Systems (KES-99)*, pp. 459–463, 1999.

[3] Joshua Grass and Shlomo Zilberstein. A value-driven system for autonomous information gathering. *Journal of Intelligent Information Systems*, Vol. 14, pp. 5–27, 2000.

[4] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[5] Arnaud Sahuguet and Fabien Azavant. Wysiwyg web wrapper factory (w4f). Available from `http://db.cis.upenn.edu/W4F/`, 1999.

[6] 市村, 長谷川, 渡部, 佐藤. テキストマイニング：事例紹介. 人工知能学会誌, Vol. 16, No. 2, pp. 192–200, 2001.

[7] 植野真臣. 意思決定アプローチによる bayesian network の因果モデル構築. 人工知能学会誌, Vol. 11, No. 5, pp. 725–734, 1996.

[8] 藤本, 賀沢, 佐藤, 島津, 北. ネット情報を使った意思決定支援 dsiu における知識獲得技術. 人工知能学会論文誌, Vol. 16, No. 1, pp. 120–129, 2001.

[9] 藤本, 山本. ネット情報を使って意思決定する過程の数理モデルの提案. 人工知能学会研究会資料 SIG-FAI/KBS-J, pp. 29–33, 2001.

[10] 富浦, 渡辺他. 特集：ここまできた自然言語処理. 情報処理学会誌, Vol. 41, No. 7, pp. 762–796, 2000.

付録：知識ベースの価値

知識ベースの価値について、「知識ベースを利用した場合に期待できる効用」に基づいた定量化を示す。

定義 1 *(知識ベースの価値)*
知識ベースとして、代替案の集合 A、情報の集合 E、確率分布 Pr が獲得されたとき、

$$V(\{A, E, \Pr\}) = \max_{a \in A} \sum_{\theta \in \Theta} U(\theta) \Pr(\theta | (E : a)) \quad (2)$$

を知識ベース $\{A, E, \Pr\}$ の価値と呼ぶ。ここに、U は効用関数、Θ は「状況」の集合、$(E : a)$ は代替案 a についての情報の集合をそれぞれ表す。　□