インターネット上の情報を扱う確率推論システムのための知識獲得エージェント

藤本 和則　　松澤 和光

日本電信電話 (株) コミュニケーション科学研究所

Abstract : In this paper, we describe a knowledge acquisition agent, which constructs probabilistic knowledge from statements written in Internet Web-pages. This agent has two functions as follows : (1) constructing probabilistic knowledge from statements written in Web-pages, and (2) gathering Web-pages from the Internet in order to use them for the knowledge construction. This paper particularly focuses on the gathering function of the agent. This gathering function has a strategy that makes agent access the Internet previously in order to decrease uncertainty in the Internet and to make gathering-plan accurately. A metric for evaluating the strategy is introduced and the strategy is evaluated by using this metric.

1　はじめに

近年、インターネットの普及により、人間は世界各国に散在する最新の情報を瞬時に取得できるようになった。こうした状況にあわせて、インターネットから有効な情報を集めて、人間の活動に役立てる研究が始められている [1]。これらの研究の多くは、ユーザの要求する情報を集めて提示する機能に注目したものであり、それらの情報に基づいた判断は、人間に委ねられている場合が多い。しかしながら、インターネットから集められた情報は、膨大であったり (大量の情報から該当するものを集めるため)、あまり知られていなかったり (次々と新しい情報へ更新されるため) することがあるので、それらから適切な判断を下す作業は困難になることが予想される。そこで、我々は、ユーザの判断に対して積極的にアドバイスを与える推論システムの実現を目指して研究を進めている [2]。

推論システムの構築には、推論に必要な知識の獲得が必須となる。我々の目指すシステムの知識については、インターネット上の情報を扱えるものであることが要求される。したがって、インターネット上に現れる最新の情報を扱えるように、知識の更新を頻繁に行う必要がある。こうした状況から、人間が直接にシステムの知識を用意するような従来の方法では、頻繁な更新に莫大なコストが見込まれ、現実的とはいえない。インターネット上の情報を扱う推論システムの実現には、こうした知識獲得のプロセスをいかにして自動化するかが重要となる。

本稿では、インターネット上の情報を扱う知識を自動的に獲得する**知識獲得エージェント**について述べる。これは、(1) 確率知識をインターネット上の記述文から構成する**構成機能**と (2) 確率知識の構成に役立つ記述文をインターネットから収集する**収集機能**の二つの機能を備えたエージェントである。インターネット上の記述文は、それ自体豊富な内容を含んでおり、また、次々に最新のものに更新されていく。したがって、もし、判断が必要となった時点で、インターネットから記述文を収集し、それらから自動的に知識を構成することができれば、インターネット上の様々な情報を扱える知識を得ることが期待できる。

知識獲得エージェントの二つの機能を説明するにあたって、ある製品 (例えばデジタルカメラ) について、ユーザが自分の好みに合った機種を選定する場面を取り上げる (図 1)。

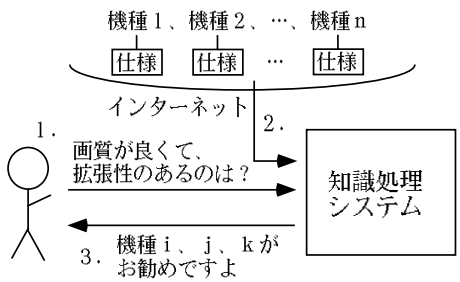

図 1: 知識処理システム

図 1において、ユーザは、デジタルカメラの性能について自分の好み (かなり画質が良く、ある程度拡張性のあるなど) を投入する (図中 1.)。また、既存のインターネット検索技術により、どのような機種が発売中か、それぞれの機種はどのような仕様か、についての情報が取得される (図中 2.)。このとき、知識処理システムは、集められた各機種について、

その仕様からそれがユーザの好みに合致するかどうかを推論し、合致するものを選んでユーザに提示する (図中 3.)。

我々は、こうした知識処理システムの推論法として統計的決定の枠組を採用した。これにより、ユーザの好みに合致する見込みを“期待効用”として定量的に算出でき、その値の大きな機種を提示するというアドバイスが可能となる。知識獲得エージェントは、この期待効用を計算するために必要となる“確率知識”を自律的に構成するものである。

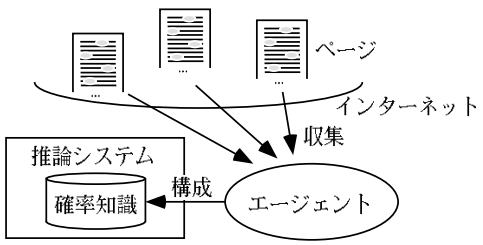

図 2: 知識獲得エージェント

知識獲得エージェントは、まず、インターネット上の記述文から確率知識を構成できなければならない。これは、知識獲得エージェントの「構成機能」により実現される。

構成機能 知識獲得エージェントは、インターネット上の記述文を「それを提供した人間のもつ主観確率[†]に制約を与える情報」として解釈する。これにより、インターネット上の記述文を「確率パラメータのとりうる数値 (確率値) の範囲を小さくする情報」として利用することが可能となる。例えば、カタログなどに記載された「仕様 s が画質を良くする」という記述は、それを記述した人間の主観確率について「仕様 s を採用すると、製品の画質が良い確率は大きくなる」という情報を与えていると解釈される。こうした情報は、次のように定式化することができる。

$$\Pr(\text{画質} = \text{良い} | \text{仕様} = s) > \Pr(\text{画質} = \text{良い})$$

さらに、知識獲得エージェントは、こうして集められた制約と確率公理を共に充足するように、確率パラメータの値を一つ決める。値を決めるにあたっては、期待誘出誤差を最小にするという基準に従う (こうした基準については 2 章に述べる)。以上のようにして、知識獲得エージェントは、インターネット上の記述文から確率知識を構成する。

2 章では、こうした知識獲得エージェントの構成機能について、「集められた制約の集合を充足するように確率知識を確定させる方法」を中心に述べる。

[†]主観確率という用語は「人間の主観から得られた確率」を「物理的な意味での確率」から区別するときに用いられる。

さて、確率知識を正確に構成するためには、確率パラメータについての制約式を多く集めて、そのとりうる値の範囲を小さいものにする必要がある。こうした制約式収集のプランニングは、知識獲得エージェントの「収集機能」により実現される。これまでに、我々は、誘出を正確にする効果の大きい制約式を優先して収集する方法を示した [2]。この手法では、確率知識を正確にする効果の大きい制約式から順に収集するので、ランダムな収集に比べ、より早く確率知識を正確にすることが期待できる。

優先順序付けに基づく収集法を実際のインターネットからの収集に適用するにあたっては、どのアドレスから、どのような制約式が得られるか分からない場合への拡張を行う必要がある。知識獲得エージェントは、どのアドレスから、どのような制約式が得られるかについて全く知らない場合には、優先順序付けを行うことができないので、ランダムにアクセスして収集するしかない。しかしながら、もし、各アドレスについて「収集できそうな制約式」を知ることができれば、そうした情報に基づいて優先順序付けを行うことができる。こうした情報は、インターネットをあらかじめ事前調査することにより獲得できるだろう。例えば、アドレスから一部のページのみをダウンロードして解析し、そのアドレスから得られるであろう制約式を予測する方法が考えられる。また、インターネット上の他のエージェントに尋ねるなどの方法 [5] も考えられる。こうした事前調査により、各アドレスについて「収集できそうな制約式」を知ることができれば、そうした情報に基づいて優先順序付けを行うことが可能となる。3 章では、まず、我々がこれまでに提案している「優先順序付けに基づく収集法」を述べ、不確実な情報環境への拡張として、**事前調査に基づく収集法**について述べる。

我々は、これまで、収集戦略の良さを定性的に論じてきたが、これを定量的に評価することができれば、より優れた収集法の実現に役立つだろう。そこで、ここでは、こうした収集法の良さを評価する基準として、*Resource-Bounded Reasoning* で研究されている枠組 [4] を導入する。これは、時間経過に伴う「答の質の増加」と「答の価値の減少」とを合わせた総合的な価値を *Utility* として定量化し、この大きさで戦略の良さを評価する枠組みである。こうした基準の導入により、時間経過に伴う答の価値の減少という考えを考慮に入れた上で、様々な収集戦略の良さを定量的に評価することが可能となる。4 章では、まず、こうした収集法の良さを評価する基準として、*Utility* の尺度について述べる。そして、ランダムな収集と、事前調査に基づく収集との *Utility* について、実際に計算機シミュレーションを行い比較を行う。

2　確率知識の誘出とその正確さ

本章では、まず、2.1節で、確率知識を誘出する問題を具体的な形で規定する。そして、2.2節で、2.1節に規定した確率知識の誘出問題のもとに、誘出の不正確さを**期待誘出誤差**として定量化する。さらに、この期待誘出誤差に基づいて、本手法により確率知識を正確に構成するための要件を論じる。なお、本章での議論は、確率制約に誤りはなく、無矛盾であるという仮定のもとに行われる。また、扱う確率変数は全て離散型とする(連続型のものは適当な量子化により離散型とできるとする)。

2.1　確率知識の誘出

我々は、集められた確率制約を充足するように確率知識を誘出する問題を次のように規定した。

定義 2.1 *(確率知識の誘出問題)*
ある確率パラメータの並びが与えられたとき、それに対応した確率値の並びを

- 集められた確率制約からくる制約
- 確率公理からくる制約 *(確率値は1を越えない非負実数など、得られた数値が"確率値"であるための制約)*

を共に充足するという条件のもとに一つ決める問題を**確率知識の誘出問題**と呼ぶ。□

定義2.1での"確率パラメータの並び"は、実際の問題では、確率値を獲得したい確率パラメータを選んで並べたものに相当する。この確率パラメータの並びをパラメータベクトルと呼ぶ。以下では、Druzdzelらの提案した分布超空間[3]の考えを導入し、確率知識の誘出問題をより具体的な問題として規定する。

Druzdzelらは、様々な確率パラメータについての制約を「標準的なパラメータ集合についての制約」として統一的に扱うため、任意の確率パラメータを導出できる"基底確率(*constituent probabilities*)"というパラメータ集合を用意した(この基底確率とは、問題領域中の全ての確率変数についての同時確率である)。そして、この基底確率によって張られる空間を"分布超空間(*distribution hyperspace*)"と呼んだ。確率制約や確率公理からくる制約は、分布超空間上の制約として統一的に扱うことができる。

例として、三つの確率変数、画質$\Theta \in \{\theta_1, \theta_2\}$、焦点方式$F \in \{f_1, f_2\}$、レンズ$L \in \{l_1, l_2\}$を取り上げる。こうした三つの確率変数からなる問題での基底確率とは、同時に起こり得る全ての場合

$$(\theta_1 f_1 l_1, \theta_1 f_1 l_2, \theta_1 f_2 l_1, \theta_1 f_2 l_2, \\ \theta_2 f_1 l_1, \theta_2 f_1 l_2, \theta_2 f_2 l_1, \theta_2 f_2 l_2) \qquad (1)$$

のそれぞれについての確率を指す(ここに、$\theta_i f_j l_k$は$\Theta{=}\theta_i, F{=}f_j, L{=}l_k$が同時に起こる事象を指す)。

ここでは、リスト(1)に対応した八つの基底確率をそれぞれ$x_1, x_2, \ldots, x_8$と書くことにする。こうした基底確率に基づいて、確率公理からくる制約は、それぞれが1以下の非負実数で、その総和が1(リスト(1)の成分は互いに排反で、全てを尽くしているため)という制約式として書くことができる。

$$0 \le x_i \le 1,\ i = 1, \ldots 8, \\ x_1 + x_2 + \cdots + x_8 = 1$$

また、確率制約からくる制約も同様に基底確率についての制約として記述することができる。例えば、

$$\Pr(\Theta = \theta_1 | F = f_1) > \Pr(\Theta = \theta_1)$$

という確率制約は、ベイズの規則と簡単な確率規則に従って、基底確率についての次のような制約式に変形して書くことができる。

$$\frac{x_1 + x_2}{x_1 + x_2 + x_5 + x_6} > x_1 + x_2 + x_3 + x_4$$

このように、確率制約や確率公理は分布超空間上の制約として統一的に扱うことができる。

確率知識の誘出問題は、この分布超空間の考えを導入して、より具体的に扱うことができる(図3)。まず、確率制約と確率公理からくる制約を分布超空間上の制約として扱い、それらを充足する分布超空間上の領域(可能領域と呼ぶ)を定める(図中 1.)。そして、可能領域中のそれぞれの点について、パラメータベクトルに対応した確率値の並び(確率値ベクトルと呼ぶ)を確率規則に従って算出する(図中2.)。この際、可能領域中の点は、可能領域の全体を反映するように選ばれる必要がある。そこで、ここでは、分布超空間の軸(基底確率)を適当なサイズに量子化して得られる量子点のうち、可能領域中にあるもの対象とする。こうして集めれた確率値ベクトルは、確率制約と確率公理からくる制約を充足するので、確率知識の誘出問題の解となりうる確率値ベクトルの集合である。ここでは、こうして集めれた確率値ベクトルの集合を**充足ベクトル集合**と呼ぶ。そして、確率知識の誘出問題をこの充足ベクトル集合から一つの確率値ベクトルを選び出す問題として扱う(図中 3.)。

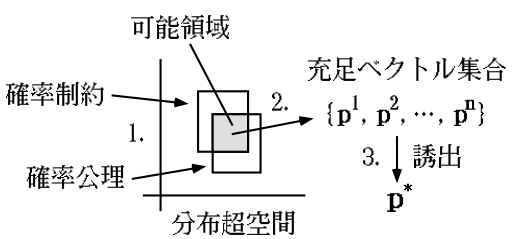

図 3: 確率知識の誘出

2.2 正確さの尺度

まずはじめに、充足ベクトル集合から一つの確率値ベクトルを誘出したときの誤差を「誘出された確率値ベクトルの“正しい確率値ベクトル”への近さ」という考えに基づいて定義する。ここでは、こうした尺度として、計算された値の意味を見通し良くするという観点から「ベクトルの各成分についての差の平均」という素朴な形のものを採用する。以下では、ここに定義した正確さの尺度を誘出誤差と呼ぶ。

定義 2.2 *(誘出誤差)*
ある n 成分のパラメータベクトルについて、誘出された確率値ベクトルを $\tilde{\boldsymbol{p}} = (\tilde{p_1}, \cdots, \tilde{p_n})$、正しい確率値ベクトルを $\boldsymbol{p} = (p_1, \cdots, p_n)$ としたとき、誘出誤差 $D(\tilde{\boldsymbol{p}} \parallel \boldsymbol{p})$ を次式で定義する。

$$D(\tilde{\boldsymbol{p}} \parallel \boldsymbol{p}) = \frac{1}{n} \sum_{i=1}^{n} |\tilde{p_i} - p_i| \qquad (2)$$

□

この誘出誤差が小さいほど、正しい確率値ベクトルに近い確率値ベクトルが得られたという意味で、より正確な誘出ができたことになる。

さて、実際には、正しい確率値ベクトルは未知なので、式(2)で定義された誘出誤差を直接知ることはできない。そこで、我々は、正しい確率値ベクトルの代わりに、2.1節で示した充足ベクトル集合を利用することを考えた。この充足ベクトル集合は「正しい確率値ベクトルとしてありうる全体」を反映すると捉えられる。したがって、我々は、誘出の正確さの尺度の一つとして、こうした充足ベクトル集合のそれぞれについて「その誘出誤差が総じて小さいように確率値ベクトルを誘出できれば正確、逆に総じて大きいようにしか誘出できなければ不正確」とする尺度を利用できるのではないかと考えた。ここでは、こうした考えに基づいた不正確さの尺度を「充足ベクトル集合の各要素についての誘出誤差の平均」という形で定義する。以下では、ここに定義した正確さの尺度を期待誘出誤差と呼ぶ。

定義 2.3 *(期待誘出誤差)*
確率制約 I のもとでの充足ベクトル集合を $\mathcal{R}_I$、誘出された確率値ベクトルを $\tilde{\boldsymbol{p}}$ $(\in \mathcal{R}_I)$ とする。このとき、期待誘出誤差 $E(\tilde{\boldsymbol{p}} \lll \mathcal{R}_I)$ を次式で定義する。

$$E(\tilde{\boldsymbol{p}} \lll \mathcal{R}_I) = \frac{1}{|\mathcal{R}_I|} \sum_{\boldsymbol{p} \in \mathcal{R}_I} D(\tilde{\boldsymbol{p}} \parallel \boldsymbol{p}) \qquad (3)$$

ここに、$|\mathcal{R}_I|$ は $\mathcal{R}_I$ の要素数を表す。 □

この期待誘出誤差が小さいほど、充足ベクトル集合のそれぞれについて、誘出誤差が総じて小さい確率値ベクトルが得られたという意味で、より正確な誘出ができたことになる。

式(3)から、期待誘出誤差は0より小さい値はとらないことがわかる(誘出誤差 D は0以上の実数値を与える関数であるため)。期待誘出誤差が0をとるのは、$\mathcal{R}_I$ の要素がただ一つの場合か、$\mathcal{R}_I$ の要素が二つ以上でそれらが全て同一の確率値ベクトルである場合のみに限られる(誘出誤差 D が0を与えるのは、二つの確率値ベクトルが同一の場合のみであるため)。これは、可能領域が小さく制限されて、ただ一通りの確率値ベクトルのみが充足ベクトル集合として許された場合、すなわち、正しい確率知識を一意に知ることができた場合のみ、期待誘出誤差は、その最小値0を与えることを意味する。それ以外、すなわち、$\mathcal{R}_I$ が二つ以上の異なる確率値ベクトルを含む場合は、期待誘出誤差は必ず0を超える値を与えることになる。また、式(3)から、期待誘出誤差は必ず1より小さい値をとることがわかる(誘出された $\tilde{\boldsymbol{p}}$ との誘出誤差が1の確率値ベクトルを前提に $\mathcal{R}_I$ の要素数を増加させていくと、式(3)の値は1に収束するが1になることはないため)。このように、期待誘出誤差は、誘出の不正確さを0以上1未満の実数値として測る尺度である。

確率知識を正確に誘出するという観点からは、期待誘出誤差の値が小さくなるように確率値ベクトルを誘出する必要がある。しかしながら、式(3)から、確率制約 I のもとでは、適切な $\tilde{\boldsymbol{p}}$ $(\in \mathcal{R}_I)$ の選択により期待誘出誤差を小さくできる限界

$$\min_{\tilde{\boldsymbol{p}} \in \mathcal{R}_I} E(\tilde{\boldsymbol{p}} \lll \mathcal{R}_I)$$

が存在することがわかる(この値は $\mathcal{R}_I$ が二つ以上の異なる確率値ベクトルを含む限り0をとらない)。これは、確率制約が集められた時点で、誘出とは無関係に、確率知識をどこまで正確に構成できるかの限界が決まってしまうことを意味する。したがって、本手法により確率知識を正確に構成するには、正確な誘出が可能なように、すなわち、“期待誘出誤差の最小値”が小さくなるように確率制約が集められていることが前提となる。

3 不確実な情報環境での収集戦略

本章では、まず、3.1節で、確率制約のもつ「誘出を正確にする効果」を**知識量**として定量化する。3.2節では、アクセスにより得られる確率制約の不確実さを確率的にモデル化し、不確実な情報環境で優先してアクセスするアドレスを与える規則を示す。そして、3.3節では、インターネットを事前調査し、その不確実さを減少させてから収集プランを立てるという収集戦略を示す。

3.1 正確にする効果の尺度

2章では、確率知識を正確に構成するには「期待誘出誤差の最小値が小さくなるように確率制約が集

められている必要がある」ことについて述べた。こうした知見に基づいて、我々は、この“期待誘出誤差の最小値”をより小さくできる確率制約の方が、誘出を正確にする効果がより大きいと言えるのではないかと考えた。ここでは、こうした考えに基づいた「誘出を正確にする効果」の尺度を「その制約を加えたときの期待誘出誤差の最小値の減少量(単純な差)」として定義する。確率制約の得られていない状態に加えた場合の変化量を知識量、すでに得られた確率制約に追加して加えたときの変化量を条件付知識量とそれぞれ呼ぶ。

定義 3.1 *(知識量と条件付知識量)*
ある確率制約 I の知識量 $K(I)$、確率制約 I_1 に確率制約 I_2 を追加したときの I_2 の条件付知識量 $K(I_2|I_1)$ をそれぞれ次式で定義する。

$$K(I) = E_*(\mathcal{R}_\phi) - E_*(\mathcal{R}_I) \tag{4}$$
$$K(I_2|I_1) = E_*(\mathcal{R}_{I_1}) - E_*(\mathcal{R}_{I_1 I_2}) \tag{5}$$

ここに、E_* は式 (3) において $\tilde{\boldsymbol{p}}$ $(\in \mathcal{R}_I)$ を動かしたときの期待誘出誤差の最小値をとる関数を表す。また、$\mathcal{R}_\phi$ は確率公理からくる制約のみから得られた充足ベクトル集合、$\mathcal{R}_{I_1 I_2}$ は I_1 と I_2 を合わせた確率制約から得られた充足ベクトル集合をそれぞれ表す。 □

この知識量、条件付知識量の値が大きいほど、期待誘出誤差の最小値をより小さくするという意味で、誘出を正確にする効果が大きいということになる。この知識量および条件付知識量の尺度を用いることにより、その値の大きい確率制約をより優先するという方法で、収集にあたっての優先順序を与えることが可能となる。

3.2 優先順序付け

本節では、まず、アクセスにより得られる確率制約の不確実さを確率的にモデル化する。そして、不確実な情報環境で、優先してアクセスするアドレスを与える規則を示す。

確率制約の全体を $\mathcal{I} = \{I_1, I_2, \ldots, I_v\}$、アドレスの全体を $\mathcal{A} = \{A_1, A_2, \ldots, A_w\}$ とする。このとき、あるアドレス A へアクセスしたとき、得られる条件付知識量の期待値は次式で与えられる。

$$V(A, \Delta) = \sum_{I_i \in \mathcal{I}} \mathrm{P}(A = I_i) \cdot K(I_i|\Delta) \tag{6}$$

ここに、確率 $\mathrm{P}(A = I_i)$ は A のデータを全てダウンロードしたとき、そこから確率制約 I_i を抽出できる確率を表し、これを**成功確率**と呼ぶ。また、Δ はすでに収集された確率制約を表す。式 (6) で与えられる尺度を**期待知識量**と呼ぶ。この期待知識量の値が大きなアドレスほど、より大きな条件付知識量を期待できるという意味で、アクセスの価値があるといえる。こうした考えに基づいて、アクセスにあたって「期待知識量の大きなアドレスをより優先する」という優先順序付けが可能となる。

不確実な情報環境での優先規則

> 確率制約の収集にあたっては、期待知識量のより大きなアドレスを優先してアクセスする。

このように、確率制約の収集のモデル化を期待知識量に基づいて行うと、インターネットの不確実さは、成功確率分布の“偏り”として捉えることができる。この“偏り”が大きいほど、期待知識量に基づいて行った優先付けの誤り率が小さくなり、より効果的な収集が期待できるようになる。

3.3 事前調査に基づく収集法

期待知識量に基づいて誤りの少ない優先付けを行うためには、成功確率分布により大きな偏りを与える必要がある。我々は、こうした「偏りの増加」は、知識獲得エージェントに、インターネット環境を自律的に事前調査する機能を持たせることにより実現できるのではないかと考えた。こうした事前調査により、成功確率分布の偏りを大きくすることができれば、より誤りの少ない優先付けを行うことが可能となる。このように、まず、事前調査によりインターネットの不確実さを減少させてから収集のプランを立てる収集戦略を**事前調査に基づく収集法**と呼ぶ。

事前調査に基づく収集法

> 1. 事前調査を実行
> 2. 成功確率の分布を更新
> 3. 各アドレスについて期待知識量を計算
> 4. 期待知識量最大のアドレスを選んでアクセス
> 5. 確率制約を収集し Δ に格納
> 6. 3. へ

事前調査に基づく収集法は、「事前調査に要するコスト」と「事前調査による偏りの増大」の二つのパラメータによって特徴付けられる。ここでは、前者を「事前調査に要する時間」、後者を「情報量」としてそれぞれ定義する (情報量の具体的な定義は付録に記す)。調査に要する時間がより小さく、そして、調査によって得られる情報量がより大きな事前調査が、事前調査に基づく収集をより効果的にするといえる。

4 収集戦略の定量的評価

本章では、まず、4.1節で、収集の戦略の良さを定量的に評価する *Utility* という尺度について述べる。4.2節では、この *Utility* の尺度を用いて、事前調査に基づく収集戦略の評価を行う。評価にあたっては、ランダムにアドレスを選んで収集する戦略の *Utility* 特性を基準とし、事前調査によりインターネットの不確実さを小さくしてからアドレスを選んで収集する戦略との *Utility* 特性との比較を行う。

4.1 評価尺度

Zilberstein らは、時間経過に伴う「答の質の増加」と「答の価値の減少」とを合わせた総合的な価値を *Utility* として定量化し、この大きさで戦略の良さを評価する枠組みを示した[4]。こうした基準の導入により、時間経過に伴う答の価値の減少という考えを考慮に入れた上で、様々な収集戦略の良さを定量的に評価することが可能となる。

Utility の時間特性 $\mathrm{U}(t)$ は、

$$\mathrm{U}(t) = \mathrm{K}(t) - \mathrm{C}(t) \tag{7}$$

として定義される。ここに $\mathrm{C}(t)$、$\mathrm{K}(t)$ は、

- コスト関数 $\mathrm{C}(t)$: 時間経過に伴う価値の減少を与える関数
- Performance Profiles $\mathrm{K}(t)$: 時間経過に伴う質 (Quality) の増加を与える関数

である。ある戦略に基づく収集法は、時刻 t_m ($t_m = \arg\max_t \mathrm{U}(t)$) に収集を終了するという規則に従って、最大値 $\mathrm{U}(t_m)$ の *Utility* を獲得することができる。したがって、ここでは、この $\mathrm{U}(t_m)$ を「その戦略により得られる *Utility* の最大値」という意味で、戦略の良さを評価する指標として用いる。

4.2 事前調査に基づく収集法の評価

事前調査に基づく収集法の評価にあたっては、まず、インターネットから確率制約を収集する過程をモデル化する必要がある。このモデルは、確率制約のモデルと成功確率のモデルとからなる。以下では、それぞれのモデルについてまとめる。なお、分布超空間の記法としては、

- 分布超空間の次元 n (n は正の整数)
- 構成割当の集合 $S = \{s_1, \ldots, s_n\}$
- 各次元の量子化サイズ $1/Q$ (Q は正の整数)

を用いた(分布超空間中の量子点数は、基底確率0を含む点を除いて Q^n とした)。

確率制約のモデル

- 形式の集合 $F = \{f_1, \ldots, f_m\}$
- 構成割当 s_i について形式 f_j で記述された確率制約を I_i^j と表す。
- 形式 f_i は i が大きくなるほど強い制約となり、確率制約 I_k^i は、確率制約 $I_k^j (i > j)$ を含意するとする。
- 確率制約 I_i^j は $\Pr(s_i)$ のとりうる範囲を $1/\mathrm{M}(f_j)$ にする。ここに M は制約の強さを数量的に与える関数 $\mathrm{M}{:}f_j \to [1, Q]$ である。

成功確率のモデル

[事前調査なし]

- あるアドレスにアクセスしたとき、内容 s_i の確率制約 I_i^* が得られる確率
  $\mathrm{P}(A = I_i^*) = 1/n,\ \forall s_i$
- あるアドレスにアクセスしたとき、形式 f_j の確率制約 I_*^j が得られる確率
  $\mathrm{P}(A = I_*^j) = 1/m,\ \forall f_j$

[事前調査あり (調査に要する時間 T)]

- あるアドレスにアクセスしたとき、内容 s_i の確率制約 I_i^* が得られる確率
  $\mathrm{P}(A = I_i^*) = p_s \ \ (1/n \le p_s \le 1)$
  s_i 以外の内容の確率制約 $I_j^* (i \neq j)$ が得られる確率
  $\mathrm{P}(A = I_j^*) = \frac{1-p_s}{n-1},\ \forall s_j (i \neq j)$
  (この確率 p_s を事前調査の**情報量指数**と呼ぶ。)
- あるアドレスにアクセスしたとき、形式 f_j の確率制約 I_*^j が得られる確率
  $\mathrm{P}(A = I_*^j) = 1/m,\ \forall f_j$

また、知識量の計算にあたっては、計算時間の削減のため、以下の近似式を用いた。

定義 4.1 *(近似知識量計算)*
各基底確率 $\Pr(s_i)$ のとりうる値を $1/q_i$ の範囲に制限にする確率制約の知識量を次式で定義する。

$$\frac{1}{|V|} \sum_{s_i \in V} \log_Q q_i$$

ここに、V は分布超空間を張る構成割当のうち、獲得する確率パラメータに関連する構成割当の集合、$|V|$ はその要素数をそれぞれ表す。□

上記モデルにおいて、分布超空間の次元数 $n = 20$、量子化サイズ $1/10$、知識量の計算対象となる次元数 $|V| = 10$ として、時刻50までの収集についてシミュレーションを行った。なお、一つのアドレスから確率制約を収集するための時間は一律1とし、確率制約の形式は $m = 10$ 種類あり、それぞれの形式 $i = 1, \ldots, m$ は、順に基底確率のとりうる範囲を $1/i$ に制限するとした。

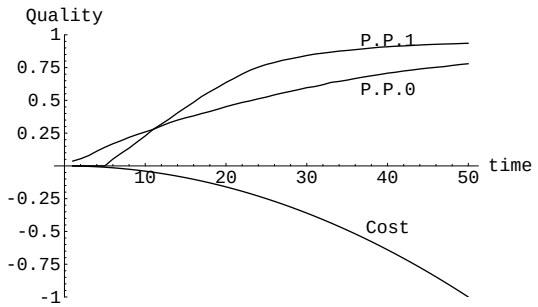

図 4: Quality の時間特性

図 4に、確率制約の収集により増加する知識量の時間特性と、時間経過に伴い知識量の価値を減少させるコストの時間特性を示す。P.P.0 は事前調査なしの収集戦略を、P.P.1 は事前調査ありの収集戦略を用いた場合の知識量をそれぞれ表す。事前調査としては、事前調査に要する時間 $T = 5$、事前調査による情報量指数 $p_s = 0.5$ の場合を取り上げた。Cost はコストの時間特性であり、これを与える関数としては、時刻0 のときに 0 をとり、時刻50 のときに -1 をとる二次関数を用いた。図 4から、得られる知識量の増加量が時間経過とともに減少していることがわかる。これは、確率制約を集めていくと、新たに集めた確率制約とそれまでに集めた確率制約との重複が大きくなっていくので、制約としての範囲限定の効果が減少するためである。なお、知識量の時間特性計算にあたっては、100 回の収集過程を試行した結果の平均をとった。

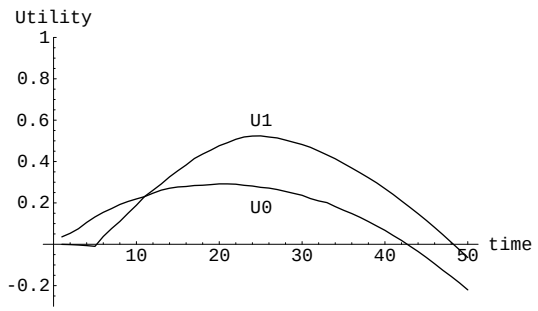

図 5: Utility の時間特性

図 5に、時間経過に伴う知識量の増加とその価値の減少を合わせた *Utility* の時間特性を示す。U0 は事前調査なしの収集戦略、U1 は事前調査ありの収集戦略の *Utility* 特性をそれぞれ表す。戦略 U0 によって得られる *Utility* の最大値は、time = 20 のとき 0.31 である。これに対して、戦略 U1 によって得られる *Utility* 最大値は、time = 25 のとき 0.53 である。このように、事前調査に基づく戦略 U1 の方が、事前調査をしない戦略 U0 より大きな *Utility* を期待できることがわかる。このように、上記のインターネット環境においては、事前調査ありの収集戦略の方がより良い戦略であるということができる。

5 おわりに

本稿では、確率知識を自動的に構成するエージェントについて、インターネットから記述文を収集する機能を中心に述べた。まず、限られた時間内に、できるだけ多くの有効な記述文を収集できるようにするため、各アドレスに「優先順序」を与え、その高い順にアクセスするという収集法を示した。この優先順序付けは、記述文の有効さの尺度である「知識量」を用いて「より大きな知識量を獲得できるアドレスにより高い優先順序を与える」という方法で行われる。この順序付けに基づく収集法は、どのアドレスにどのような記述文があるかが分かっていることを前提とするが、実際のインターネット環境では、この仮定は妥当でない場合が多い。そこで、本稿では、インターネットを事前調査することにより、「どのアドレスにどのような記述文があるか」についての不確実さを減らしてから収集を開始するという戦略を示した。さらに、こうした収集の戦略を定量的に評価する枠組を導入し、提案の収集法の有効性について論じた。今後は、本稿で取り上げた機種選択問題を対象に、実データを扱いながら研究を進めていく予定である。

参 考 文 献

[1] 武田 英明, “ネットワークを利用した知的情報統合”, 人工知能学会誌, Vol. 11, No. 5, pp. 680–688, 1996.

[2] 藤本 和則, 松澤 和光, “インターネット上の記述文から確率知識を構成する一手法 —構成の基本原理を中心に —”, 1998 年情報学シンポジウム, pp. 129–136, 1998.

[3] Marek J. Druzdzel and Linda C. van der Gaag, “Elicitation of Probabilities for Belief Networks: Combinaiton Qualitative and Quantitative Information,” *Proceedings Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 141–148, 1995.

[4] Shlomo Zilberstein and Victor Lesser, “Intelligent Information Gathering Using Decision Models,” *UMASS Technical Report 96-35*, 1996.

[5] Tim Oates, Nagendra Prasad, and Victor Lesser, “Cooperative Information Gathering: A Distributed Problem Solving Approach,” *UMASS Technical Report 94-66*, 1994.

付録: 事前調査により得られる情報量の定義式

あるアドレス A についての事前調査 $(\mathrm{P}_0 \rightarrow \mathrm{P}_1)$ により得られる情報量は次式で与えられる。

$$J_A(\mathrm{P}_1) = \mathrm{H}(\mathrm{P}_1) - \mathrm{H}(\mathrm{P}_0)$$

ここに、H はアドレス A のエントロピーを与える関数で、次式で与えられる。

$$H_A(\mathrm{P}) = -\sum_{I_i} \mathrm{P}(A = I_i) \log \mathrm{P}(A = I_i)$$