

サイバー空間と実空間の統合的情報検索

山田誠二 間瀬心博

東京工業大学大学院総合理工学研究科知能システム科学専攻

〒 226-8502 神奈川県横浜市緑区長津田町 4259

yamada@ymd.dis.titech.ac.jp

あらまじ WWW のように有用であり公開されている情報がある一方で、有用であるが非公開で、さらには、明文化されておらず人間の頭の中にだけある情報も多数存在する。本研究では、クエリを与えれば、後は WWW とユーザグループの両方から、公開情報と非公開情報の両方を効率よく検索する統合的な情報検索システム HERIS (HEterogeneous Resource Information Search) を提案する。

キーワード 情報検索、公開/非公開情報、マルチエージェントシステム、WWW

Integrated information search in the WWW and a human group

Seiji Yamada Motohiro Mase

CISS, IGSSE, Tokyo Institute of Technology

4259, Nagatsuta, Midori-ku, Yokohama 226-8503, Japan

yamada@ymd.dis.titech.ac.jp

Abstract

In this paper, we propose a framework for searching information through both the WWW and a human group. Though the information retrieval using a search engine in the WWW is very useful, we can not acquire local information owned by person and not explicitly described in text. User knows neither where target information is in the WWW nor who knows in a human group. Thus we integrate the information retrieval in WWW with that in a human group, develop heterogeneous resource information search HERIS as multi-agent system.

key words

Information retrieval, open/closed information, multi-agent system, WWW

1 はじめに

現在、インターネット、特に WWW (the World Wide Web) を情報源とした情報の獲得が急速に広まりつつある。WWWには、現在数億の規模のWebページが世界中に分散して存在しており、時々刻々更新されているため、情報源として非常に有用である。しかし、膨大な情報が無秩序に存在するため、ユーザが欲しい情報の所在を知ることは難しく、検索エンジンを用いて所望の情報を探す情報検索が一般的に行われている。

一方、有用な情報であるにも関わらず、一般に公開されていない情報もたくさん存在する。例えば、LANの設定やアプリケーションソフトウェアのインストールなどに関する Tips は、他者にとっても広く有用な情報である場合が多いが、その情報のかなりの部分は、明文化されていないか、あるいは、明文化されていても一般に公開されていない。我々は、前者の一般に公開されている情報を公開情報、後者のような公開されない情報を非公開情報と呼ぶ。

このように有用である非公開情報にアクセスするもつとも有効かつ効率的な方法の一つは、その情報を所有している人に直接たずねることである。しかし、どの人にたずねればよいかという情報源の選択の問題、つまり情報検索が、公開情報と同様に必要となる。さらに、重要なことは、検索するユーザは、自分の欲しい情報が、公開情報なのか非公開情報なのかを自分で判断して、Webの検索エンジンにクエリを渡すか、その情報を知っているであろう人に問い合わせるか、あるいはその両方をしなければならない。しかし、このような作業は煩雑であり、クエリを与えれば、後は公開情報と非公開情報の両方を効率よく検索して、その結果を統合して分かりやすく提示してくれるシステムが望まれる。本研究では、我々はそのような検索システムとして、WWWとユーザグループの両方から目的の情報を検索し、複数の情報源が見つかった場合には、アクセスすべき情報源を絞り込んでユーザに提示す統合的な情報検索システム HERIS(HEterogeneous Resource Information Search)を提案する。

HERISは、ユーザに割り当てられているユーザエージェントとWWWの検索エンジンに割り当てられる検索エンジンエージェントから構成されるマルチエージェントシステムとして構成され、エージェント間の通信プロトコルとして契約ネットプロトコル [1] を用いる。ユーザがユーザエージェントにクエリを与えると、システムは、他の情報源、つまり他のユーザエージェントや検索エンジンエージェントにクエリをタスクを告知し、検索結果と情報源の属性からなる入札を集め。そして、それらの入札の情報を基に、有用な情報源を統合してユーザに提示する。検索主体であるユーザが選択した情報源が人間であれば、音声と動画による直

接対話を確立し、Webページ等公開情報であればブラウザを用いて情報を提示する。

ブックマークファイルに蓄積されたWebページに関する個人の知識を共有する研究では、協調フィルタリングが利用した森らによるブックマークエージェント [5] がある。このシステムはブックマークされているURLから取り出したHTMLファイルを解析し、プロファイルを構築するエージェントにより構成されるマルチエージェントシステムである。ブックマークエージェントは、担当ユーザのブックマークを監視し、必要に応じて他のエージェントと通信することでユーザの現在見ているWebページと類似したページを提示する。しかし、ブックマークエージェントは、ブックマークなど限定された情報を共有するに止まり、ユーザ間の直接対話を支援する枠組みにはなっていない。

また、仮想空間内でのユーザ間のコミュニケーションに関する研究では、FreeWalk [6] がある。FreeWalkはコンピュータネットワーク内でのユーザ間の簡単な対話の機会を提供している。しかし、このような研究では、WWWとユーザグループの両者を検索する統合的な枠組みは用意されていない。

2 情報のアクセス容易性と明文化

2.1 情報の分類の軸

公開情報、非公開情報の概念は、情報へのアクセス容易性 (accessibility) という軸で説明できる。明文化されている情報は、公開/非公開として二分できるが、明文化されていない情報、つまり人間の頭の中にだけある情報は、人にたずねるしかなく、そのとき、情報源である人にアクセスできるか否かの他に、その人の状態によって連続的にアクセス容易性が変化すると考えられる。例えば、忙しいときは、問い合わせたとしても、答えが返ってくる可能性が低いという意味でアクセスが困難になる。よって、アクセス可能性という連続な軸を考えることにより、情報源が人である場合も、明文化された文書である場合も扱うことができる。

このように、アクセス容易性という軸で情報を分類して考えるとともに、情報を分類するもう一つの軸として、明文化/非明文化 (explicit/implicit) を導入する。これは、情報が明文化されているか否かの二律背反である。現実には、明文化の困難さやセキュリティの問題から、明文化されずに、人間の頭の中だけに存在する情報は多く存在するが、そのような非明文化情報を明文化情報を変換していく過程は、本研究のように知識の共有を目指す場合に重要であるため、この軸を導入する。

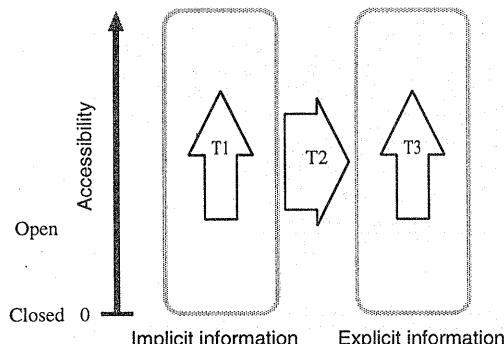


図 1 情報の状態と遷移

2.2 情報の状態遷移

以上の 2 つの軸を考えた結果、情報の分類とその遷移は、図 1 のように表現できる。ここで、情報には、図のように T1, T2, T3 つの遷移が可能であり、その遷移の意味は、以下のようなになる。

- 遷移 T1：非明文化情報のアクセスが、より容易になる遷移。
- 遷移 T2：非明文化情報が、明文化され、明文化情報になる遷移。
- 遷移 T3：明文化情報のアクセスが、より容易になる遷移。

一般的に、情報は、明文化され、かつアクセスが用意である方向、つまり、図 1 の右上の方向に向かうべきであると考えられるので、T1, T2, T3 の 3 つの情報の遷移を支援するシステムは、有用である。本研究では、後述するように、遷移 T1, T3 を主に扱うが、将来的には、T2 も射程にいれる予定である。

3 HERIS システム

3.1 システムの概要

HERIS システムの概要を、図 2 に示す。HERIS は、WWW と人間のグループを統合的に検索するシステムであり、ユーザエージェントと検索エンジンエージェントからなるマルチエージェントシステムとして構成される。エージェント間の通信には、契約ネットプロトコル [1][8] を用いる。ユーザと検索エンジンのそれぞれには、ユーザエージェントと検索エンジンエージェントが割り当てられ、それぞれが担当の情報源に関して、ユーザプロファイルと検索エンジンプロファイルを管理する。各プロファイルは、知識プロファイルと資源プロファイルの 2 つのサブプロファイルで構成されている。

知識プロファイルは、ユーザや検索エンジンが持っている情報あるいは情報に関連した単語とその重みで構成され、資源プロファイルは、ユーザの実空間での状態や検索エンジンの負荷などの、情報源について知識以外の情報を管理する。

ユーザがクエリをシステムに与えると、システムは契約ネットプロトコルを通して適切な情報源を選定する。ユーザには、適切な情報源から得られた Web ページのヒットリスト、および音声・動画を用いて直接情報を提供してくれるユーザのリストが提示される。

以下に、HERIS の情報検索の手続きの概要を説明する。

1. ユーザが HERIS の自分を担当しているユーザエージェントにクエリを与えると、そのユーザエージェントはマネージャとなり、他のすべてのエージェントは契約者となる。
2. マネージャは、すべての契約者に対してクエリを含んだタスク告示をブロードキャストする。
3. この告示を受けた契約者は、自分の担当するユーザ、検索エンジンの知識プロファイルとクエリの適合度を計算し、計算した適合度と資源プロファイルの値を合わせて入札メッセージとしてマネージャに返信する。
4. マネージャは、受け取った入札メッセージを検討して適切な契約者を選択し、その契約者にクエリを依頼メッセージとして送信して、検索結果が含まれた結果メッセージを受け取る。
5. マネージャは、受け取ったメッセージを統合し、Web ページのヒットリストや直接音声・動画を通じて情報を提供してくれるユーザのリストを提示する。
6. ユーザは、その提示された情報源から、適切なものを選択し、そこからクエリに適合した情報を得る。

3.2 エージェントの契約手続き

ユーザエージェントと検索エンジンエージェントの詳細な手続きは、以下のようになる。まず、大別してマネージャモードと契約者モードの 2 種類のモードがある。クエリを受け取るのは、ユーザエージェントだけなので、検索エンジンエージェントは、常に契約者モードで動作する。また、ユーザエージェントと検索エンジンエージェントの手続きは、入札メッセージで扱うプロファイルが異なるだけで、他は同じである。

マネージャモード ユーザエージェントが、ユーザからクエリを受け取るとマネージャとなる。

1. タスク告示メッセージを他の全てのエージェントにブロードキャストする。

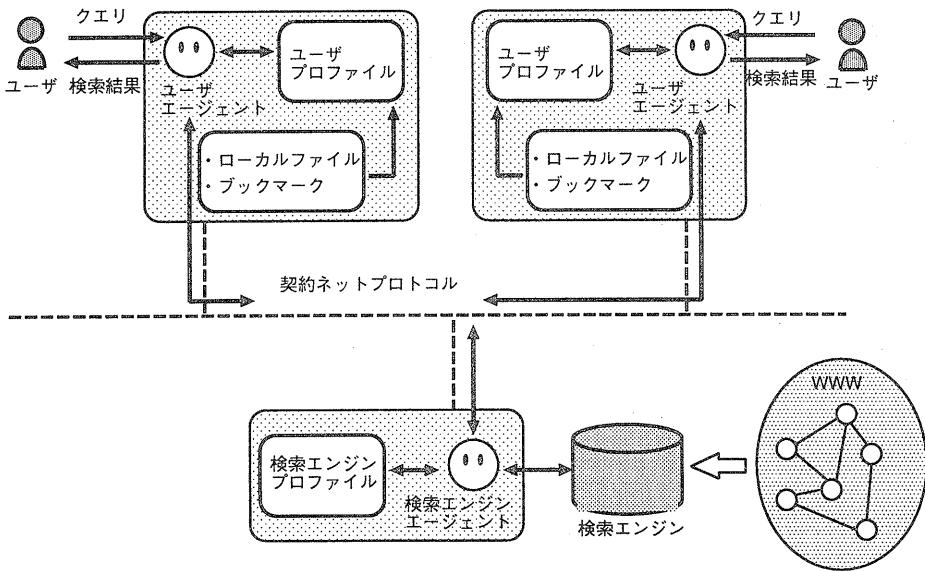


図2 システムの概要

2. 契約者からの入札メッセージを受け取る。
3. 入札メッセージを検討し、適切な契約者を選定する。選定した契約者エージェントに依頼メッセージを送り、契約を結ぶ。
4. 契約者からの結果メッセージを受信する。
5. 検索結果を統合してユーザに提示する。

契約者モード マネージャからタスク告示メッセージを受け取ると、そのタスクに対しては、契約者となり、契約者モードで動作する。

1. タスク告示メッセージ内のクエリと、担当ユーザ、あるいは担当検索エンジンの知識プロファイルとの適合度を計算することにより、クエリと担当情報源の適合度を求める。具体的な計算方法は、後述する。
2. 適合度が閾値をこえた場合、入札メッセージをマネージャに送信することにより、入札を行なう。
3. マネージャから依頼メッセージを受けた場合には、現在の適合率、検索結果を結果メッセージとしてマネージャに返信する。

3.3 契約のメッセージ

エージェント間の契約手続きで用いるメッセージは、タスク告示メッセージ、入札メッセージ、依頼メッセージ、結果メッセージの4種類である。また、各メッセージには、基本的に、タイプ、ラベル、送信元、送信先

の4種のスロットが用意されており、さらにメッセージによってスロットが追加される。4種類のメッセージと、そのスロットおよびスロット値の例を以下に示す。

タスク告示メッセージ

- タイプ：タスク告示
- ラベル：契約1
- 送信元：ユーザエージェント1
- 送信先：全てのエージェント
- タスク内容：情報検索
- クエリ：[cscw, www, information_retrieval]
- 入札期限：2000年7月18日14時58分

入札メッセージ

- タイプ：入札
- ラベル：契約1
- 送信元：検索エンジンエージェント2
- 送信先：ユーザエージェント1
- 適合度：クエリと知識プロファイルとの類似度
- 資源プロファイル

依頼メッセージ

- タイプ：依頼
- ラベル：契約1
- 送信元：ユーザエージェント1
- 送信先：検索エンジンエージェント2
- タスク内容：情報検索
- クエリ：[cscw, www, information_retrieval]

結果メッセージ

- タイプ：結果
- ラベル：契約 1
- 送信元：検索エンジンエージェント 2
- 送信先：ユーザエージェント 1
- 結果：

検索エンジンの場合

- ヒットリスト：適合率の高い Web ページのリスト

ユーザの場合

- ユーザ名
- IP アドレス
- ポート番号：vic, vat で使用するポート番号。

4 プロファイルの構成

ユーザプロファイルと検索エンジンプロファイルの構成を以下に示す。両プロファイルとも、知識プロファイルと資源プロファイルからなる。なお、知識プロファイルは、ステミングされた単語を次元として、その重みを値とする単位ベクトルで表現される。このような知識プロファイルの表現形式は、ユーザプロファイルと検索エンジンプロファイルに共通したものである。

4.1 ユーザプロファイル

知識プロファイル 文書集合 $D = \{d_1, \dots, d_N\}$ が与えられたときの文書ベクトルの生成方法を以下に示す。TF/IDF と文書構造を用いた重み付けを併用している。この手続きは、ユーザプロファイルの知識プロファイル生成に共通したものである。

1. TF/IDF を用いて、下式で文書 d_i における単語 t_i の重み $w_{ij} = TF_{ij} \times IDF_j$ を計算する。ただし、 D の要素である文書 d_i における語 t_j の出現回数を f_{ij} 、 D 中で t_j が出現する文書数を n_j とする。

$$TF_{ij} = \frac{f_{ij}}{\sum_j f_{ij}}$$

$$IDF_j = \log \left(\frac{\max_j n_j}{n_j} \right) + 1$$

2. 上で決まった w_{ij} に対し、文書が構造化されている場合、つまり、LaTeX のソースファイルや HTML ファイルの場合は、そのタグ構造を用いてさらに重みを追加する。すなわち、<TITLE>, <Hn>, \bf, \title, \sectionなどには含まれた単語の重みを大きくする。

この知識プロファイルは、ユーザが持っている非公開情報を扱う。その非公開情報は、明文化情報、非明

文化情報からなり、それぞれの獲得、管理方法を以下に示す。

- ユーザの明文化情報：WWW には公開されていない、ユーザの所有する文書ファイルに書かれている情報のことである。これらの文書ファイル内で高い頻度で出現する単語は、ユーザがよく知っている情報に関連する単語と考えられるので、ユーザのホームディレクトリ下の文書ファイルの集合を文書集合 D として、上記の手続きにより、ベクトル V_1 を生成する。
- ユーザの非明文化情報：この情報は、明文化されていない、ユーザが知っているだけの状態の情報である。ユーザが普段よく見ている Web ページにはユーザが興味を持ち、よく知っている情報があると考えられため、ユーザの Web ブラウジングを監視したり、ブックマークされている URL から HTML 文書を集め、それを文書集合 D として、上記の手続きにより、ベクトル V_2 を構成する。さらに、より現実的な方法として、ユーザから知っている情報に関連のあるキーワードを明示的に入力してもらい、それを基に、 V_2 を修正する。

ベクトル V_1 と V_2 を、まず単純に合成し、重みでソートした上位一定数の次元を知識プロファイル V_K とする。

知識プロファイルは、システムの初回起動時に、ユーザの指定する範囲のローカルな文書ファイルやブックマークされている Web ページから単語を抽出し作成する。そして、それ以降は、基本的にユーザに手動で更新してもらう。

資源プロファイル 資源プロファイルは、管理対象がユーザ、検索エンジンによって保持する内容が違う。ユーザの資源プロファイルでは、在席可能性、認知的負荷、社会的関係について管理する。

- 在席可能性：ユーザエージェントは、CCD カメラから得られる画像を用いた簡単な画像認識により、ユーザが現在在席しているか否かの確信度である在席可能性を計算する。 $[0, 1]$ の範囲の値をとる。
- 認知的負荷：ユーザエージェントは、ユーザのエディタ、Web ブラウザなどの日常的によく使用するアプリケーションの使用頻度を監視して、認知的負荷を計算する。つまり、使用頻度が高ければユーザが作業に集中していると考えられるため、認知的負荷を大きくして、他のユーザからの音声・動画での直接対話を繋がりにくくすることができる。 $[0, 1]$ の範囲の値をとる。

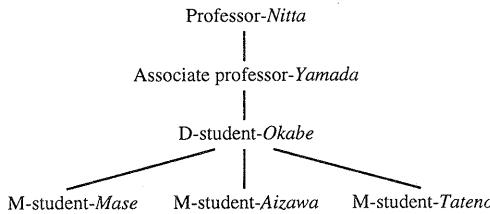


図3 ユーザグループの社会的地位のオントロジー

- 社会的関係：ユーザエージェントは、ユーザの所属しているグループ内でのユーザ間のオントロジーを用いて社会的関係を計算する。もし契約者エージェントが担当するユーザが、マネージャが担当するユーザよりもグループ内の高い地位にいる場合には、この値は負の値をとる。反対の場合には正の値をとる。マネージャが高い地位にいるユーザを担当し、契約者が低い地位にいるユーザを担当している場合、音声・動画での直接対話が繋がりやすくなる。社会的関係は、 $[-1, 1]$ の範囲の値をとる。オントロジーの例を、図3に示す。

4.2 検索エンジンプロファイル

知識プロファイル MetaWeaver[4], SavvySearch[2]等で、用いられている統計手法を利用して、検索エンジンプロファイルを構築する。基本的な手法は、サンプルとなるクエリを担当検索エンジンに与えて、そのヒットリストを解析することにより、どのクエリに対し、担当検索エンジンがどの程度適合したWebページを返せるかを評価して、知識プロファイルを構成する。

資源プロファイル 下記のネットワーク負荷を、検索エンジンの資源プロファイルとして持つ。

- ネットワーク負荷：ネットワークの込み具合いを示す。検索エンジンに ping コマンドを送り、反応時間を測定することにより評価する。反応時間が長ければ、この値は高くなる。 $[0, 1]$ の値をとる。

5 適切な情報源の選択

5.1 クエリとプロファイルの適合度の計算

§3.2 の契約者モード手続きの1で示した、契約者がタスク告知メッセージ中のクエリと担当情報源の適合度を、その知識プロファイルを基に計算する方法について述べる。この手法は、契約者が、ユーザエージェントでも、検索エンジンエージェントでも同じである。

基本的には、知識プロファイルとクエリをベクトルとして扱い、その内積を計算して、適合度をする。この方法は、情報検索の一般的に用いられているベクトル空間モデル[7]である。まず、クエリをステミングした単語を次元とした単位ベクトル V_q で表す。前述のように、知識プロファイル V_K は、ステミングされた単語とその重みからなる単位ベクトルなので、 V_q と V_K の内積をとることにより、適合度が計算できる。

計算した適合度がある閾値 τ を越えた場合に、契約者は、マネージャに対して入札メッセージを送信する。

しかし、この手法ではクエリとプロファイル内の単語が厳密に一致しない限り、適合度は0になり、ほとんどヒットしないことが考えられるため、実用的ではない。よって、概念辞書を用いてクエリの類義語を利用したマッチングを用いるか、またはユーザに対してクエリに関する情報の有無の問い合わせを直接行なうことを考えている。

5.2 適切な契約者エージェントの選択方法

§3.2 のマネージャモードの手続き3で、マネージャは、適切な契約者を選択するが、その方法を以下に示す。マネージャは契約者から受け取った入札メッセージ内の情報から、ユーザと検索エンジンそれぞれに対し、その情報源の重要度 E_u, E_s を求める。

そして、求めた重要度の上位数個のエージェントに対して、依頼メッセージを送信する。重要度 E_u, E_s は、下式を用いて求める。

$$E_u = w_r * \text{適合度} + w_p * \text{在席可能性} \\ + w_s * \text{社会的関係} - w_c * \text{認知的負荷}$$

$$E_s = w_r * \text{適合度} - w_n * \text{ネットワーク負荷}$$

この重要度は、通常の情報検索のように適合度を評価するだけでなく、それに加えてユーザの在席、社会的関係などの実際的に重要な特性も統合した評価であることに注目されたい。

現時点では閾値 τ 、重み w_r, w_p, w_s, w_c, w_n は、実験的に決めているが、今後はユーザ毎に適切な値を学習する方法を考えている。

6 検索結果の統合と提示

ユーザが得られる情報は、他のユーザのローカルファイル、明文化されていない情報、LAN内で公開されているWebページと検索エンジンの提示するヒットリストから得られるWebページなどである。

このうちユーザの持つ非公開情報は、その情報を知っているユーザ本人に直接尋ねるしかない。そのため、これら的情報がヒットした場合には、情報を持っているユーザ名が最終的な検索結果として提示され、検索主体のユーザが希望し、かつ相手のユーザが了解すれば、

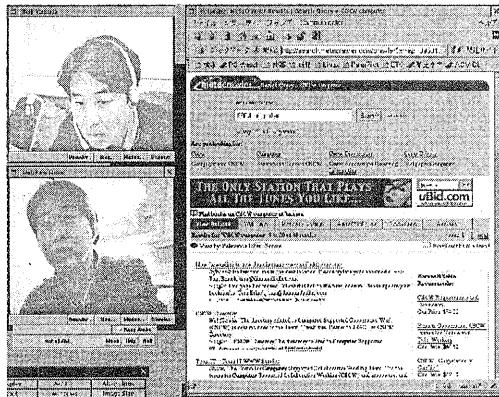


図 4 統合された検索結果（イメージ図）

ネットワークを介した音声・動画による直接対話アプリケーションを接続して、情報を得ることができる。また、公開されている Web ページはユーザが直接参照することができるので、その URL が Web 検索エンジンのヒットリストと同様の検索結果として提示される。図 4 に、結果表示のイメージ図を示す。

7 実装

現在、C, Ruby, GTK+などのプログラミング言語を用いて HERIS を構築中で、ほぼ完成の状態である。ユーザ間直接対話を実現するための音声・動画通信には、それぞれ vat¹, vic²[3] を用いている。

8 情報の遷移との対応付け

本節では、§2.2 で述べた情報の遷移と HERIS の機能との関係を説明する。以下のような対応が考えられる。

- 遷移 T1: HERIS では、vic, vatなどを用いたネットワーク経由の対話の実現、ユーザの状態（在席可能性、認知的負荷）を考慮した情報源の評価により、ユーザへの直接の問い合わせをより容易にしている。これらの機能が、非明文化情報のアクセスがより容易になる遷移 T1 に対応する。
- 遷移 T2: 非明文化情報が、明文化され、明文化情報になる遷移 T2 は、現在のところ、HERIS では実現されていない。
- 遷移 T3: HERIS は、ホームディレクトリの下の文書ファイルから、ユーザの知識プロファイルを自動生成する機能をもつ。これにより、明

¹<http://www-nrg.ee.lbl.gov/vat/>

²<http://www-nrg.ee.lbl.gov/vic/>

文化情報のアクセスが、より容易になる遷移 T3 を実現している。

現在実現されていない遷移 T2 については、今後アクセスの多い非公開かつ非明文化情報については、その情報を所有するユーザに、明文化、つまり文書にまとめるなどを推奨するエージェントの開発を考えている。また、同じような推奨エージェントとして、既に明文化されているが、非公開である情報を WWW に公開することを進めるエージェントの開発により、遷移 T3 の更なる実現が可能になると考える。

9 まとめ

本稿では、マルチエージェントシステムの枠組を用いて、WWW とユーザグループを通じて公開/非公開、明文化/非明文化情報を統合的に検索するシステム HERIS を提案した。情報源である、各ユーザと各検索エンジンに、エージェントが割り当てられ、契約ネットプロトコルによる通信を行なう。

また、各エージェントが、担当ユーザ、担当検索エンジン毎に作成するプロファイルを利用することで、システムは適切な情報源から選択的に情報検索が可能である。ユーザプロファイルと検索エンジンプロファイルは、知識プロファイル、資源プロファイルから構成され、情報の適合度だけではなく、ユーザの状態や検索エンジンのつながり易さなどの実際的に重要な情報源の特性も考慮した評価により、適切な情報源の選択が行なえる点を特徴とする。HERIS の利用により、検索主体であるユーザは、目的の情報の所在、性質を気にかけることなく、情報検索することが可能になる。

今後は、システム全体の実装を完了し、HERIS の情報検索の能力を実験的に評価する予定である。

謝辞

本研究に関して議論していただいた山田研究室のメンバーに感謝いたします。

参考文献

- [1] R. Davis and R. G. Smith. Negotiation as a metaphor for distributed problem solving. *Artificial Intelligence*, Vol. 20, No. 1, pp. 63–109, 1983.
- [2] A. E. Howe and D. Dreilinger. Savvy search: a metasearch engine that learns which search engines to query. *AI Magazine*, Vol. 18, No. 2, pp. 19–25, 1997.
- [3] S. McCanne and V. Jacobson. vic: a flexible framework for packet video. In *Proceedings of*

the third international conference on Multimedia,
pp. 511–522, 1995.

- [4] M. Mori and S. Yamada. Adjusting to specialties of search engines using metaweaver. In *WebNet 2000 – World Conference on the WWW and Internet*, 2000. to appear.
- [5] 森幹彦, 山田誠二. ブックマークエージェント：ブックマークの共有による情報検索の支援. 電子情報通信学会論文誌 D-I, Vol. J83-D-I, No. 5, pp. 487–494, 2000.
- [6] H. Nakanishi, C. Yoshida, T. Nishimura, and Toru Ishida. FreeWalk: Supporting casual meetings in a network. In *International Conference on Computer Supported Cooperative Work (CSCW-96)*, pp. 308–314, 1996.
- [7] G. Salton. *Automatic Text Processing*. Addison-Wesley, 1989.
- [8] R. G. Smith. The contract net protocol: High-level communication and control in a distributed problem solver. *IEEE Transaction on Computer*, Vol. 29, No. 12, pp. 1104–1113, 1980.