

# インターネット上での知的情報統合

北 村 泰 彦†

インターネット上には現在膨大な量の情報が蓄積されているが、ブラウザによる現在の情報検索手法には様々な問題がある。本稿はこれらの問題に対処する知的情報統合技術を情報検索、情報組織化、情報配送の観点から解説する。

## Intelligent Information Integration on the Internet

YASUHIKO KITAMURA†

Although a huge amount of information is stored on the Internet, current information retrieve methods mainly by using browsers have several drawbacks. This paper surveys intelligent information integration methods to deal with the drawbacks from viewpoints of information retrieval, organizing, and distribution.

### 1. はじめに

いまやインターネットは現代社会の中に急速に浸透しており、われわれの日常生活を支えるインフラストラクチャの一つとなってきた<sup>10)</sup>。インターネットが提供するサービスの中でも利用者間での情報共有を実現するWWW(World Wide Web)は最も人気の高いものである。WWWがこれほどまで多くの利用者を獲得した原因としては、以下のものがあげられる。

- 世界中に分散して存在する様々な情報源がURL(Universal Resource Locator)の指定により統一的にアクセス可能となる。
- 情報提供者はHTML(Hyper Text Markup Language)により記述されたWWWページをサーバ上に作成することで全世界に向けて容易に情報発信を行なうことが可能である。
- WWWページにはテキストだけでなく、画像、音声、動画などを埋め込むことができ、そのレイアウトの自由度も高い。またハイパーリンクにより分散して存在するページを相互に関連付けることができる。
- WWWページはブラウザ(browser)を介して閲覧することができるが、マウスのみでほとんどの操作を行なうことができ、計算機の初心者であって

も容易に使いこなすことができる。

ネットワーク上で情報共有あるいは情報処理を行なうことはそれほど目新しいことではないが、WWWによる情報共有が従来のそれと大きく異なっている点は、不特定多数の利用者を前提としている点であり、このような特徴はビジネスネットワーク、研究教育ネットワーク、社会共同体ネットワークなど、21世紀における地球規模でのさまざまなネットワークアプリケーションを生み出す源となることが期待される。一方で、このような不特定利用者を前提とした「開かれたネットワーク」は多くの技術的課題を同時に抱えている。WWW技術の問題点をまとめると以下ようになる。

- (1) 自律分散サーバ：情報は複数のサーバに分散して存在しており、それぞれ個別に維持管理されている。したがって利用者はどこにどのような情報があるか分からない。
  - (2) 大量の情報：アクセス可能な情報量は膨大であり、利用者が必要な情報のみを検索することは容易でない。
  - (3) あいまい・不統一な情報構造：提供される情報に統一な構造がないので、機械的な情報処理が困難である。また提供されている情報が冗長であったり、誤りが含まれていたりすることもある。
  - (4) 動的な情報源：情報が頻繁に更新されるサーバがあるが、利用者はその変更気づかない場合がある。
- 以上のような問題点のゆえに、利用者が自ら必要としている情報を適切に検索することすら困難を覚える

† 大阪市立大学工学部  
Faculty of Engineering, Osaka City University

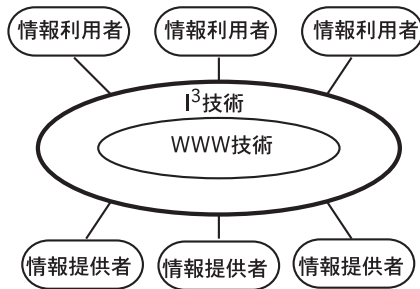


図1 知的情報統合 (I<sup>3</sup>) 技術による情報共有

ことが多い。また分散している関連情報をまとめあげたり、更新された情報をタイムリーに入手するにも多くの困難が伴うのが現状である。このような問題点に対して、分散データベースをはじめとする旧来の分散処理技術は十分に対応することができない。なぜならこれらの技術は、たとえ情報が複数のサーバに分散して存在していたとしても、仮想的には集中的な管理手法を適用することが可能であるという「閉じたネットワーク」を前提としているからである。

そこで本解説では「開かれたネットワーク」を前提とした新しい情報共有技術について紹介をおこなう。そしてこれらの技術の総称としてインターネット上での知的情報統合 (Intelligent Information Integration: I<sup>3</sup>) 技術と呼ぶことにする。

## 2. 知的情報統合技術の分類

知的情報統合技術は第1図にも示すように、WWW技術をベースとし、情報提供者 (information provider) と情報利用者 (information consumer) 間での情報共有の促進、改善、効率化を行なう技術である。

I<sup>3</sup>技術は大きく以下の三つの分野に分類することができる<sup>21)</sup>。

- 情報検索：情報利用者のWWW情報源アクセスを支援する。情報利用者が必要とする情報を含むWWWページのURLを知らない場合に、それを示したり、それが多数である場合には絞込みを行なう。これは前節で述べた課題の(1),(2)への対処である。
- 情報組織化：WWW情報源から提供される情報を組織化し、その付加価値を高める。これは前節で述べた課題(3)への対処である。
- 情報配送：情報提供者と情報利用者間の情報の流れを改善、促進する。これは前節で述べた課題(4)への対処である。

本解説では以下、それぞれの分野における研究事例や実用化事例を紹介する。

## 3. 情報検索

WWW情報源はインターネット上に分散しており、それぞれのサーバ上で個別に情報提供者により管理されている。利用者はURLによりWWWページを最小要素としたアクセスが可能であるが、それを知らなければアクセスすることは難しい。ここでは情報利用者がインターネット上から必要な情報を探し出すことを支援する情報検索手法について紹介する。

### 3.1 検索エンジン

この問題に対処した代表的なサービスが検索エンジン (search engine) である。検索エンジンはYahoo! に代表されるインデックス系とAltaVista やgoo に代表されるロボット系の二種類に大きく分類される。インデックス系検索エンジンは人手によるWWWページの閲覧や情報提供者からの要望に応じて、何らかのシソーラスに従ってディレクトリを構成するものである。ロボット系検索エンジンはソフトウェアロボットによりWWWページを自動的に収集することで、ディレクトリを構成するものである。これらの二つの方式にはそれぞれ長所短所がある。インデックス系ではシソーラスに応じて分類がされているので必要な情報を検索することが容易であるが、情報収集は人手によっているのでその情報量は少ないという短所がある。ロボット系では大量の情報が利用可能である一方で、分類は不十分であるという短所がある。ロボット系検索エンジンにおける情報検索の手段としてはキーワード検索が主流であるが、キーワードによっては利用者の意図とのミスマッチを起こすことが多い。例えば、“buffalo”というキーワードで検索した場合、検索エンジンにはそれが動物を指すのか、地名を指すのかが区別できないという問題がある。またロボット系検索エンジンでは格納されている情報が大量であるので、必要以上の検索結果が返されるという問題がある。例えば先述のgooで“buffalo”を用いて検索すれば9268個ものURLが返され(1998年10月23日現在)、それらの内容を全て確認することは実質的に不可能である。このような問題に対処するために、TITAN<sup>4</sup>では検索結果を分類して表示するインタフェースを提供している。またHusky Search<sup>5</sup>は検索エンジンの結果に含まれる内容記述(オリジナルWWWページの

<http://www.yahoo.com/>

<http://www.altavista.digital.com/>

<http://www.goo.ne.jp/>

<sup>4</sup> <http://titan.mcnet.ne.jp/index-x2.html>

<sup>5</sup> <http://zhadum.cs.washington.edu/zamir/cluster.html>

断片)をもとに二次的な分類を行っている<sup>23)</sup>。

### 3.2 メタ検索エンジン

検索エンジンの問題点に対処しようとする試みの一つにメタ検索エンジン (meta search engine) がある。メタ検索エンジンは既存の検索エンジンに対して並行に検索を行ない、その結果を統合して、検索結果の質を向上させようとするものである。大手の検索エンジンであっても収集している情報には偏りがあるのでメタ検索エンジンは一つの有効な手段であるといえる<sup>7)</sup>。代表的なメタ検索エンジンには MetaCrawler があり、Lycos, Infoseek, WebCrawler, Excite, AltaVista, Yahoo などの汎用検索エンジンを統合している<sup>18)</sup>。MetaCrawler では以下のようにして検索結果を統合している。まず、各検索エンジンから得られた URL のスコアを 1000 点満点に換算する。次に、複数の検索エンジンから同じ URL が得られた場合には重複を取り除き、全てのスコアの和をその URL のスコアとする。最後にスコア順に URL をソートする。このように多くの検索エンジンの検索結果に含まれた URL では全体的に高いスコアとなる。

しかしながら MetaCrawler では一つの検索要求に対して多数の検索エンジンを並行して動作させるために多くの計算資源や通信資源を浪費してしまうという問題も生じる。SavvySearch では検索キーワードの種類や資源の状況に応じて動的に利用する検索エンジンの組み合わせを変化させるような並列検索戦略を取っている<sup>8)</sup>。この場合、キーワード  $q$  に対する検索エンジン  $s$  のランク付けをするためにまず以下に示すメタインデックス  $Q_{q,s}$  を計算している。

$$Q_{q,s} = \sum_{t \in q} M_{t,s} \times I_t / \sqrt{T_s}$$

ここで  $M_{t,s}$  は検索エンジン  $s$  に関する単語  $t$  のメタインデックス重みで、検索エンジンが結果を返さなければ  $-1$  点、返した結果に対して利用者が  $n$  個のリンクをたどれば  $n$  点付加される。 $I_t$  は単語  $t$  の出現頻度の逆数で頻出語を排除する役割を果たしている。また  $T_s$  は検索エンジン  $s$  のすべてのメタインデックスの絶対値の和であり、汎用エンジンより専門エンジンを重視する役割を果たしている。さらに動的に変化するネットワークの負荷や検索エンジンの状況を考慮して最終的なランクは

$$R_{q,s} = Q_{q,s} - (P_{s,h} + P_{s,r})$$

で与えられる。ここで  $P_{s,h}$  と  $P_{s,r}$  は検索エンジン  $s$

のヒット率と通信遅延に関するペナルティで検索エンジンがアクセス不能になったり、通信遅延が大きくなるとそのペナルティも大きくなる。SavvySearch ではこのようなランク付けを動的に学習しながら、利用する検索エンジンの数と組み合わせを決定している。

一般に検索エンジンは検索キーワード入力の方法、検索結果の表示法などがそれぞれまちまちであり、メタ検索エンジンの構築を難しくしている。そこで検索エンジンの統合をより容易にするために、STARTS<sup>6)</sup> と呼ばれる検索プロトコルの標準化に対する試みも行われている。

### 3.3 専門検索エンジン

検索エンジンを高度化させるもう一つの方法はその専門化である。最近では企業や大学などの一組織であっても大量の WWW 情報を所有している場合があり、その検索を支援するために組織内の情報に特化した検索エンジンは多く見られる。またニュースやショッピングなど特定の用途に特化された検索エンジンものもある。この中で興味深いものの一つにインターネット上でのショッピングを支援する Jango がある。Jango では購入したい商品を指定すると検索ロボットが自動的に起動され、複数のショッピングサイトを検索し、価格順にソートして表示してくれる。

このような専門検索エンジンを開発するには多くの部分をハードコーディングに頼る必要があるが、検索エンジンの検索結果をフィルタリングすることにより専門化したものとして Ahoy! があげられる。Ahoy!<sup>4)</sup> は複数の情報源と利用者からの入力情報をもとに個人のホームページを探し出す専用検索エンジンである<sup>19)</sup>。Ahoy! が利用する情報源には汎用検索エンジン (前述の MetaCrawler)、電子メールアドレスデータベースなどを用い、利用者は探したい個人の名前、組織、メールアドレス、国名などを入力する。Ahoy! は基礎となるデータとしては検索エンジンの結果を用い、大量の出力結果を組織の名前やメールアドレスを手がかりにふるいにかけるという dynamic reference sifting という手法を用いている。例えば、探したい個人のメールアドレスが kitamura@info.eng.osaka-cu.ac.jp であったとしよう。このとき検索エンジンから得られる多量の URL の中でドメイン名に info.eng.osaka-cu.ac.jp を含むものを選択し、そこからさらに絞り込みをかければ良いことが分かる。このほか WWW ページのタイトルや内容をもとに個人ホームページかどうかの絞り込みを

<http://www.metacrawler.com/>  
<http://guaraldi.cs.colostate.edu:2000/form>

<http://jango.excite.com>  
<sup>4</sup> <http://ahoy.cs.washington.edu:6060/>

行なっている。このような手法は個人ホームページに特化した知識やヒューリスティックを用いているので他の用途に直接利用することはできないが、既存の汎用検索エンジンに比較して情報発見の正確さ（誤った結果を返さない性能）や再現性（存在するならば結果に含まれる性能）に優っている。

### 3.4 ブラウジング支援

WWWにおける情報検索の方法としては検索エンジンなどの助けをかりて直接URLを指定して目的とするWWWページに至る方法と、あるWWWページから始めていろいろなハイパーリンクを辿りながら目的とするWWWページに至る方法がある。前者は利用者が情報検索に関してははっきりとした目的をもつ場合であり、後者は利用者が漠然とした目標しか持たない場合に取られる方法である。

後者における利用者支援としては、利用者が指定した興味を示すキーワードと過去のブラウジング履歴をもとに、ブラウジングをナビゲートするWebWatcherがあげられる<sup>11)</sup>。WebWatcherでは利用者の興味とリンクを関連付けるために以下のような目標関数を用いている。

$$\text{LinkQuality} : \text{Page} \times \text{Interest} \times \text{Link} \rightarrow [0, 1]$$

ここでPageは現在ブラウジング中のページ、Interestは利用者の興味、Linkは推薦の候補となるリンクであり、それぞれ単語ベクトルとして表現され、TFIDFヒューリスティックで重み付けされた内積を取ることにより評価される。そして評価の高いリンクはWWWページ上にハイライトされたり、動的に付加されたりする。

さてWebWatcherではこの目標関数を学習するためにブラウジング履歴を用いている。すなわち、ブラウジングにおいてある利用者がリンクを選択するとその利用者の興味キーワードが注釈としてそのリンクの単語ベクトルに付加される。この手法によりリンク推薦の性能が向上することが示されている。また単純なリンクの推薦だけでなく、強化学習の手法を用いて適切なWWWページのナビゲーションも支援している。

以上のような利用者個人の明示的なフィードバックを元にWWWブラウジング支援に関する研究事例としてはWebWatcher以外にもLetizia<sup>16)</sup>やSyskill & Webert<sup>1)</sup>などがある。一方、不特定多数の利用者の

WWWアクセスの履歴を利用することでブラウジング支援を行うシステムとしてはfootprints<sup>22)</sup>があげられる。WWWシステムにおけるリンクはWWWページの設計者が自由に設定することができるので、それはあたかもジャングルのように張り巡らされている。しかしながらそのリンクをたどる利用者のパターンにはある程度の規則性が存在するはずである。有用なリンクに対しては多くの利用者がたどるであろうし、そうでないリンクをたどる利用者はわずかであろう。footprintsではWWWサーバに対する不特定多数の利用者からのアクセスログを記録し、そのパターンをJavaAppletを用いて表示する。利用者は利用頻度の高いリンクを知ることができるので、ブラウジングの手助けとなる。このパターンはブラウジング支援だけでなく、WWWページの設計に対しても有用である。

footprintsでは特定のサーバに関するブラウジング支援を行うだけであったが、この手法をさらに大規模にし、商用化したシステムとしてAlexa<sup>23)</sup>がある。Alexaをインストールすると画面上にツールバーのようなものが表れ、現在ブラウジングを行っているWWWページのアクセス頻度や人気などの情報と、次にたどるべきサイトの情報などがAlexaのサーバから受信し、表示する。このツールバーの動きは利用者のブラウジングを支援するだけでなく、ブラウジングの履歴を逆にAlexaのサーバに送り返している。このようないわゆるgive and takeの手段で大量のブラウジング履歴を入手し、それをブラウジング支援にフィードバックしている。

## 4. 情報組織化

インターネット上に分散して存在するWWWサーバ上には有用な情報を提供しているものが数多く存在する。これらのサーバはブラウザを介して個別に運用されていることが多いが、相互運用が可能になれば、多くの付加価値を生み出すことが期待される。例えば、鉄道会社、航空会社、ホテルなどのWWWサーバを組み合わせることができれば旅行の計画や予約などを一元的に行なうことが可能になるであろう。本節では、インターネット上に存在するさまざまな情報源を統合し、統一的な視点を与えたり、相互運用を可能にする情報組織化を支援するための試みを紹介しよう。

### 4.1 異種情報源の統合

米国Stanford大学のTsimmisプロジェクトはデー

---

TFIDFヒューリスティックは以下の式で与えられる。

$$TFIDF(w, d) = TF(w, d) \times \log n / DF(w)$$

ここで $TF(w, d)$ は単語ベクトル $d$ に現れる単語 $w$ の回数、 $DF(w)$ は単語 $w$ を含む単語集合の数、 $n$ は単語集合の総数である。この式により単語集合と特徴的な単語との関係が得られる。

---

<http://footprints.media.mit.edu/>

<http://www.alexa.com/>

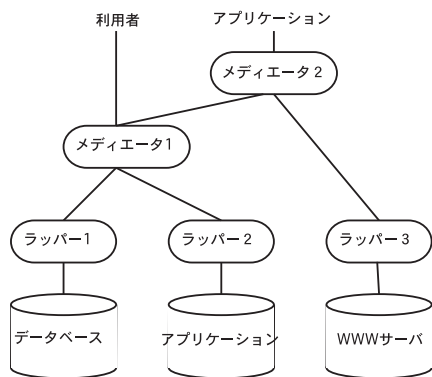


図2 Tsimmisアーキテクチャ

```

<図書リスト, set, {図書1, 図書2}>
図書1 : <図書, set, {著者1, タイトル1, 価格1}>
  著者1 : <著者, set, {著者11, 著者12}>
    著者11 : <著者名, str, "Ullman">
    著者12 : <著者名, str, "Widom">
  タイトル1 : <タイトル, str, "Databases">
  価格1 : <価格$, float, 58.00>
図書2 : <図書, set, {著者2, タイトル2, 価格2}>
  著者2 : <著者, str, "増永">
  タイトル2 : <タイトル, str, "データベース">
  価格2 : <価格¥, int, 2400>

```

図3 OEMでの記述例

データベースに限らずネットワーク上に分散している様々な異種情報源を柔軟に統合することを目的としている<sup>17)</sup>。このプロジェクトの特徴は従来の分散データベースでとられたように情報源のレベルで統合するのではなく、それぞれの自立的情報源の出力結果をメディアータ(mediator)とよばれるモジュールにより統合する点である。

Tsimmisアーキテクチャを第2図に示す。データベース、アプリケーションソフトウェア、WWWサーバなどの異種情報源はラッパー(wrapper)によって情報源固有のインタフェースから共通情報モデルへと変換される。Tsimmisプロジェクトではこの共通情報モデルとしてOEM(Object Exchange Model)が用いられている。OEMではデータオブジェクトの入れ子を許すとともにデータの意味を表すラベルがつけられている。例えば文献リストは第3図のように表現される。

この例に示されるようにOEMではデータの構造(著者)や型(価格)が情報源ごとに異なっても統一的に扱うことができる。またメディアータはOEMデータを変換したり、組み合わせたりすることでアプリケーションや利用者が望むビューを提供する働きを持って

いる。

以上のような異種情報源を統合する手法としてはTsimmisプロジェクトのようなデータベースからのアプローチだけでなく、CORBAやHORBなどの分散オブジェクトからのアプローチや次節で述べる連邦アーキテクチャ<sup>5)</sup>やRETSINA<sup>20)</sup>のようなマルチエージェントシステムからのアプローチもある。

#### 4.2 エージェント通信言語

Tsimmisプロジェクトにおいて情報統合の中心的役割を果たすメディアータは基本的には手作業により開発される。連邦アーキテクチャでは異種情報源をエージェント化し、それらの相互作用による論理ベースの情報統合を実現しようとしている。連邦アーキテクチャにおいてエージェントを結びつける中核的技術がエージェント通信言語(ACL: Agent Communication language)である<sup>4)</sup>。

エージェント通信言語は三つのレベルから成り立っている。最初のレベルはエージェントが利用する語彙である。エージェントが通信するためには、対象物やその関係を指し示す何らかのこばを用いなければならない。そしてそのこばの利用に関しては一貫性がないなければならない。この一貫性を保つ一つの方法はアプリケーション領域にオントロジと呼ばれる共通の辞書を作ることであり、エージェント開発者はこの辞書を参照しながらプログラムを行う。もちろん人間の世界にも様々な言語があるようにこのようなオントロジは一つとは限らない。そのような場合にはオントロジの変換を行うような仕組みも必要になる。

第二はエージェント間で交換するメッセージの内容を記述するレベルであり、このための記述言語として一階述語論理を拡張したKIF(Knowledge Interchange Format)が採用されている。例えば、KIFでは社員番号、所属、給与といった単純な事実は

```
(salary 015-46-3946 widgets 72000)
```

のように表現する。

第三のレベルはメッセージ通信に関するエージェントの意図を記述するもので、KQML(Knowledge Query and Manipulation Language)が採用されている。例えば、同一のメッセージ内容(produce 200 widgets)であっても以下に示すように意図の違いによりその意味は異なってくる。

- (ask-if (produce 200 widgets)) : 送信者は受信者に述べられている事実が真かどうか尋ねている。「部品を200個生産しますか?」

- (tell (produce 200 widgets)) : 送信者は受信者に彼の信念を通信している。「部品を200個生産します。」
- (achieve (produce 200 widgets)) : 送信者は受信者にその事実が真となるように要求している。「部品を200個生産しなさい。」
- (deny (produce 200 widgets)) : 送信者は受信者に述べられている事実が真ではないことを通信している。「部品を200個生産しません。」

KQMLでは言語行為理論に基づく上に示す ask-if や tell といった30種あまりの遂行語 (performative) を用いてエージェント間の相互作用をサポートしている。なお以上のようなエージェント通信言語は FI-PA (Foundation for Intelligent Physical Agents) において標準化が進められている。

#### 4.3 情報抽出

Tsimmis プロジェクトでは wrapping により、また協調アーキテクチャではエージェント化することにより異種の情報源に共通のインタフェースを与え、統合可能にしていた。WWW 情報源は通常のデータベースのようなデータの構造化が行われていないので、その中から必要な情報を抽出することは容易ではない。しかし WWW 情報源からの情報抽出を支援するプログラミング言語の研究に関するいくつかの事例がある。

WebL はブラウジング操作を記述する service combinator と WWW ページからの情報抽出を記述する markup algebra から構成されるオブジェクト指向のスクリプト言語である<sup>12)</sup>。

例えば、www.xxx.com のアクセスに失敗した場合はそのミラーである www.mirror-xxx.com にアクセスし、その結果を変数 page に保存することは以下のように記述される。

```
page := getpage(" http://www.xxx.com ") ? getpage(" http://www.mirror-xxx.com ")
```

ここで?は逐次実行を表している。これ以外にも並列実行や繰り返しなどのコマンドが定義されている。また page の中で、UCI という文字列を含む全てのレベル2 ヘッダーを変数 titles に入れるコマンドは以下のようなになる。

```
titles := page.Elem("H2") contain page.Pat("UCI")
```

その他の研究としては MetaCommander があり、ゲノムデータベースの統合に応用されている<sup>13)</sup>。

<http://www.csel.toronto.edu/fipa/>  
<http://www.research.digital.com/SRC/WebL/>  
<http://www.kdel.info.eng.osaka-cu.ac.jp/MetaCommander/>

また WebSQL<sup>4)</sup> は WWW 情報源を関係データベースと見なし、ページレベルの情報検索が可能な問い合わせ言語である<sup>2)</sup>。さらにプログラミング言語以外の研究としてはテンプレートを用いて情報抽出する試みもある<sup>9)</sup>。

#### 5. 情報配送

WWW 技術が広く社会に浸透した一因には WWW ブラウザの利便性があげられるが、これは利用者が URL で指定された WWW サーバの情報を引き出すという一方向的なものである。したがって通常のブラウザ利用法ではアクセスするまでサーバ側の情報が更新されていることが気づかないという問題点がある。最近では、ニュースや株価情報などリアルタイムで情報更新される WWW サーバが増えており、利用者側からも最新の情報を即座に入手したいという要求も高くなっている。また情報提供者の視点からも、ビジネスや商品などの情報は、利用者からのアクセスを受身で待つのではなく、それを能動的に送り届けたいという要求がある。そこでインターネット上に存在する情報を情報提供者から情報利用者に対してスムーズに配送するための技術が重要になる。

##### 5.1 ブル技術とプッシュ技術

従来のブラウザによる情報アクセス技術は WWW サーバから利用者が情報を引き出すという意味でプル技術と呼ばれる。このようなプル機能を自動化するための仕組みとして、Internet Explorer における購読機能や、オートパイロットプログラムなどがある。これらは利用者が指定した WWW ページを定期的に巡回し、その変更を知らせたり、自動的にダウンロードしてオフラインでのブラウジングを可能にするものである。

プル技術とは逆に、情報提供者から能動的に情報配送するプッシュ技術としては、Internet Explorer におけるチャンネル機能や PointCast<sup>5)</sup> などの商用システムがすでに運用されている。これらのシステムは情報提供を行なうことを目的としているが、Marimba 社<sup>6)</sup> の Castanet では Java で記述されたアプリケーションプログラムを配送することができ、パソコン上でアプリケーションソフトウェアの自動バージョンアップなどの用途にも利用することが可能である<sup>15)</sup>。

##### 5.2 情報マッチメイキング

インターネットを介して利用可能な情報は急激に増

<sup>4)</sup> <http://www.cs.toronto.edu/~websql/>

<sup>5)</sup> <http://www.pointcast.com/>

<sup>6)</sup> <http://www.marimba.com/>

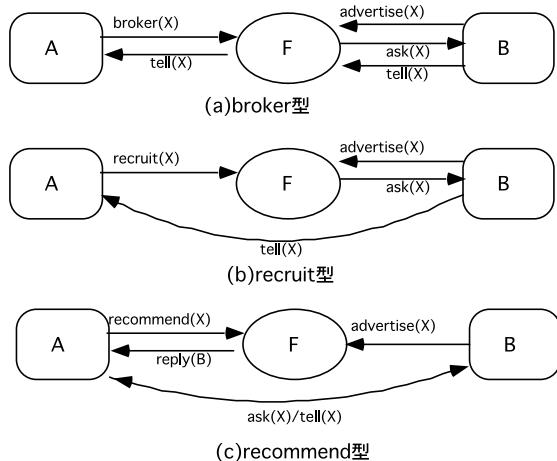


図4 協調促進器を介したマッチメイキング

加しており、従来の情報検索手法では対処できないほど、情報源の数が増え、動的に変化しているので、利用者は自動的な通知サービスを必要としている。情報マッチメイキングは協調アーキテクチャをベースにしており、情報提供者が提供可能な情報の仕様を、情報利用者が必要な情報の仕様を協調促進器 (facilitator) に送ることで自動的な情報配送を行うものである<sup>14)</sup>。

情報マッチメイキングの形態は図4に示すように三つの型がある。

- (a) broker 型: エージェント A は X に関する情報が必要なことを broker(X) を用いて知らせ、エージェント B は X に関する情報を保持していることを advertise(X) で知らせる。そこで協調促進器が B に問い合わせを行い、その結果を A に転送する。
- (b) recruit 型: エージェント A は X に関する情報が必要なことを recruit(X) を用いて知らせ、エージェント B は X に関する情報を保持していることを advertise(X) で知らせる。そこで協調促進器が B に問い合わせを行うが、情報は協調促進器を経由せずに直接 A に送られる。
- (c) recommend 型: エージェント A は X に関する情報が必要なことを recommend(X) を用いて知らせ、エージェント B は X に関する情報を保持していることを advertise(X) で知らせる。協調促進器はエージェント A にエージェント B のアドレスを知らせ、情報の問い合わせと転送はエージェント A とエージェント B の間で行われる。

### 5.3 リコメンダーシステム

インターネット上での情報流通を議論する際には、プル技術、プッシュ技術といったような情報流通の方法論だけでなく、その流れをどのように制御するかと

いうことも重要になる。すなわち情報提供者と情報利用者の興味的一致に応じて適切に情報を配送することが要求される。ここでは電子商取引の分野においてその利用が期待されているリコメンダー (推薦) システムを紹介しよう<sup>3)</sup>。

コンピュータを用いた推薦の方法には大きく分けて二つある。一つは内容ベース推薦でキーワードなどを用いて直接商品を推薦する方法である。内容ベース推薦の原型は文書検索システムにあり、利用者の指定するキーワードに対してシステムはそれに関連する重要なキーワードを含む文書を推薦する。しかしながらこの手法は、映画、音楽、レストランなどキーワードでは表すことのできない感性が決め手となる商品の推薦には適していない。

このような欠点を克服するもう一つの手法として協調的フィルタリング (collaborative filtering) が注目されている。これは利用者の商品購入履歴を手がかりに、音楽 CD などの個人の嗜好を強く反映するような商品の推薦を行なうものである。例えば、顧客 X が商品 A と商品 B を購入し、顧客 Y が商品 A を購入したとしよう。この時、システムは顧客 X と同一の商品 A を買った顧客 Y は顧客 X と嗜好が同様であるという仮定を行ない、顧客 Y のまだ購入していない商品 B の推薦を行なう。もちろんこの手法では購入履歴情報が少ない場合や、新製品を推薦するような場合には有効ではないが、大量の購入履歴を電子的に入手可能な電子商取引のような分野には有望な手法であると考えられている。リコメンダーシステムを採用した商用システムとしては Firefly が代表的である。

## 6. ま と め

インターネット上での新たな情報共有を支援する知的情報統合技術について、情報検索、情報組織化、情報配送の観点からの事例を紹介した。多くのものは脚注に示した URL よりアクセス可能となっているので、実際に利用されることをお勧めする。しかしながら、これらの技術は満たすべき要求に対して完全にこたえているわけではなく、研究すべき技術的課題は多く残されている。知的情報統合技術は人工知能をはじめとし、ネットワーク、分散処理、文書検索、データベース、ユーザインタフェースなどの多くの要素技術が融合されたものであり、それぞれの研究分野の垣根を越えた学際的研究領域であるといえる。インターネット

社会フィルタリング (social filtering) と呼ばれることもある。  
<http://www.firefly.net/>

はそのような研究者の協力を支援するものであると同時に、その実験場にもなりうるものであるといえる。

## 謝 辞

本解説の内容に関して議論いただいた「けいはんな知的情報統合研究会」の皆さんに感謝します。

## 参 考 文 献

- 1) Ackerman, M. et al.: Learning Probabilistic User Profiles, *AI Magazine*, Vol. 18, No. 2, pp. 47-56 (1997).
- 2) Arocena, G.O., Mendelzon, A.O., and Mihaila, G.A.: Applications of a Web Query Language, *Proc. of 6th WWW Conf.* (1997). <http://atlanta.cs.nchu.edu.tw/www/PAPER267.html>
- 3) Balabanovic, M. and Shoham, Y.: Fab: content-based, collaborative recommendation, *Communications of the ACM*, Vol. 40, No. 3, pp. 66-72 (1997).
- 4) Finin, T., Labrou, Y., and Mayfield, J.: KQML as an Agent Communication Language, *Software Agents*, AAAI Press, pp.291-316 (1997).
- 5) Genesereth, M.R. and Ketchpel, S.P.: Software Agents, *Communications of the ACM*, Vol. 37, No. 7, pp. 48-53(1994). <http://logic.stanford.edu/sharing/papers/agents.ps>
- 6) Gravano, L., Chang, K., Garcia-Molina, H., Lagoze, C., and Paepcke, A.: STARTS: Stanford Protocol Proposal for Internet Retrieval and Search, <http://www-db.stanford.edu/~gravano/start.html> (1997).
- 7) 林良彦, 小橋喜嗣: WWW上の検索サービスの技術動向, 情報処理, Vol. 39, No. 9, pp. 861-865 (1998).
- 8) Howe, A.E. and Dreilinger, D.: Savvy Search: a metasearch engine that learns which search engines to query; *AI Magazine*, Vol. 18, No. 2, pp. 19-25 (1997).
- 9) Hsu, J.Y. and Yih, W.: Template-based information mining from HTML documents, *Proceedings of 14th National Conference on Artificial Intelligence*, pp. 256-262 (1997).
- 10) 石田晴久: インターネットはここまで来た-現状と課題-, 情報処理, Vol. 39, No. 5, pp. 393-399 (1998).
- 11) Joachims, T., Freitag, D. and Mitchell, T.: WebWatcher: a tour guide for the world wide web; *Proceedings of 15th International Joint Conference on Artificial Intelligence*, pp. 770-775 (1997). <http://www.cs.cmu.edu/afs/cs/project/theo-6/web-agent/www/ijcai97.ps>
- 12) Kistler, T. and Marais, H.: WebL - a programming language for the Web, *Proc. of 7th WWW Conf.* (1998). [http://www.research.digital.com/SRC/personal/Johannes\\_Marais/pub/www7/paper.html](http://www.research.digital.com/SRC/personal/Johannes_Marais/pub/www7/paper.html)
- 13) 北村泰彦, 野崎哲也, 辰巳昭治: スクリプトに基づくWWW 情報統合支援システムとゲノムデータベースへの応用, 電子情報通信学会誌, Vol. J81-D-I, No. 5, pp. 451-459 (1998).
- 14) Kuokka, D. and Harada, L.: Integrating Information via Matchmaking, *Journal of Intelligent Information Systems*, Vol.6, pp.261-279 (1996).
- 15) L. リメイ: Castanet インターネット環境でのプログラム配布と公開, プレゼンティスホール(1997).
- 16) Lieberman, H.: Letizia: an agent that assists web browsing; *Proceedings of 14th International Joint Conference on Artificial Intelligence*, pp. 924-929 (1995). <http://lieber.www.media.mit.edu/people/lieber/Lieberary/Letizia/Letizia.html>
- 17) 宗像浩一: 異種分散情報源の統合, システム/制御/情報, Vol. 40, No. 12, pp. 514-521 (1996).
- 18) Selberg, E. and Etzioni, O.: The MetaCrawler architecture for resource aggregation on the web; *IEEE Expert*, Vol. 12, No. 1, pp. 11-14 (1997). [papers/ieee/ieee-metacrawler/ieee-metacrawler.html](http://papers.ieee/ieee-metacrawler/ieee-metacrawler.html)
- 19) Shakes, J., Langheinrich, M. and Etzioni, O.: Dynamic reference sifting: a case study in the homepage domain; *Proceedings of 6th International World Wide Web Conference* (1997). <http://www.cs.washington.edu/homes/jshakes/ahoy-paper/paper.html>
- 20) K. Sycara, Pannu, A., Williamson, M., Zeng, D. and Decker, K.: Distributed Intelligent Agents, *IEEE Expert*, Vol.11, No.6 pp.36-45 (1996).
- 21) 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). 武田氏がまとめた <http://ai-www.aist-nara.ac.jp/doc/people/takeda/iii/index.html> も参考になる。
- 22) Wexelblat, A. and Maes, P.: Footprints: Visualizing Histories for Web Browsing. <http://wex.www.media.mit.edu/people/wex/Footprints/footprints1.html>
- 23) Zamir, O. and Etzioni, O.: Web Document Clustering: A Feasibility Demonstration, *Proc. SIGIR98* (1998). <http://zhadum.cs.washington.edu/zamir/sigir98.ps>