マルチメディア通信と分散処理 80－19
グループウェア 21－19
（1997. 1. 31）

分散情報探索のための情報管理エージェント

阿部康一 武田利浩 丹野州宣
E-mail: {kouichi, takeda, tanno}@etn.yz.yamagata-u.ac.jp
山形大学工学部電子情報工学科

近年、インターネットと総称されるコンピュータ・ネットワークの普及が急速に展開され、その利用者は企業や学術研究機関のみならず一般家庭にまで広がって来ている。とりわけ、誰もが自由に情報の発信者となることができるという理由で、World-Wide Web (WWW) の利用者は益々増加の傾向にある。それにより、瞬く間にインターネット上に膨大な情報が無秩序に氾濫・蓄積されるに至った。このような状況の中、利用者が自分の必要とする情報をどのように見つけ出すかが大きな問題となってきている。そのため現在では、World-Wide Web 上の情報を対象とした様々な検索システムが盛んに研究・開発され続けている。しかしながら、既存の検索システムのほとんどが単一ホストによる集中管理方式を採用しており、インターネット上の情報資源全体を管理するには限界がある。

そこで本研究では、World-Wide Web などの情報資源を提供するサーバ自身が検索サーバとしての役割をもつ分散情報の自己管理方式を提案する。本方式では、エージェントのマイグレーションによって分散配置された情報資源を相互に結び付け情報探索に利用する。これにより、インターネット上の情報資源を効率的に管理し、利用者の必要とする情報を探索することが可能になると考えている。本文では、システムの基本構造について述べ各情報資源を管理する情報管理エージェントに焦点を当て、その情報管理方式について詳細に説明する。

An Information Management Agent for Distributed Information Retrieval

Kouichi ABE, Toshihiro Taketa, Kuninobu Tanno
E-mail: {kouichi, takeda, tanno}@etn.yz.yamagata-u.ac.jp
Department of Electrical and Information Engineering, Faculty of Engineering, Yamagata University

The current exponential growth of the Internet precipitates the increase and the dispersion of information. All sorts of information retrieval systems for the World-Wide Web have been studied and developed in the world. But their systems have the limits of the information management.

The purpose of our research is to build an efficient information sharing by means of an original distributed information management and retrieval methodology. In this paper, we suggest a prototype system using multi-agent model, and describe its basic structure and functions. We describe the information management agent from the point of view of the information retrieval as well.

1 はじめに

近年、インターネットと総称されるコンピュータ・ネットワークの普及が急速に展開され、その利用者は企業や学術研究機関のみならず一般家庭にまで広がって来ている。とりわけ、誰もが自由に情報の発信者となることができるという理由で、World-Wide Web (WWW) の利用者は益々増加の傾向にある。それにより、瞬く間にインターネット上に膨大な情報が無秩序に氾濫・蓄積されるに至った。

このような状況の中、現在、利用者が自分の必要としている、あるいは興味のある情報をどのように見つけ出すかが大きな問題となってきている。最近では、膨大な World-Wide Web という情報資源から目的の情報を見つけ出すために多種多様の検索システムが研究・開発され続けている。一般的な検索システムでは、これらの情報を管理する方法として情報カタログを利用している。この方式では、あらかじめ管理対象となる情報に関するデータベース (情報カタログ) を作成しておき、ユーザからの検索要求を満たす情報を情報カタログから検索して、その結果の一覧をユーザに提示する。

しかしこの場合、ユーザが必要な情報をインターネット上から見つけるには、検索システムのサービスを行っているサーバ、あるいはそこへのリンクなどを明確に知っている必要がある。また現在の World-Wide Web やネットニュースなどの情報量を見てもわかるように、現在のように単一ホストで情報資源全体を把握していくには限界がある。さらに、日々新しい World-Wide Web のサーバが誕生し続ける中、その新たなサーバの情報までも管理対象に即座に追加するのは困難である。このように、インターネット上の情報資源の極度の分散化、大規模化が現在の情報検索における大きな問題となってきている。

そこで本研究では、従来の単一ホストによる集中型の情報管理方式に対して、複数のホストによる分散型の自己情報管理方式を導入した分散情報共有システムを提案する。このような分散型自己情報管理方式においては各情報資源間の効率的な検索、つまり相互に情報を結びつける方法が必要である。本研究ではマルチエージェント・モデル [1] の概念を導入し、その自律的なエージェントのマイグレーション機能を利用する。さらに、遺伝的アルゴリズムの概念にもとづくエージェントの自律再生産と最適な検索エージェントへの進化を介して、分散環境上における情報の効率的な検索を実現する。本研究の究極的な目標は、World-Wide Web などに限らずインターネット上の全ての情報資源の効率的な情報共有の実現である。

以下、まず 2章で本研究で提案するシステムの基本概念について説明する。次に 3章で本システムにおける情報管理エージェントの機能とその情報管理方式について説明し、最後に 4章でまとめと課題を述べる。

2 分散情報共有システム ATRAS

2.1 情報管理・検索方式

現在の情報検索システムのほとんどが、単一ホストで検索サービスの対象として限定された情報資源の情報をデータベース化して管理している (図 2.1)。大部分の検索システムにおける検索方法としては、ユーザが直接検索サーバに接続して検索を行う方式である。この時、専用のクライアント (ユーザ・インタフェース) が必要となることもある。

また、これらの検索システムの情報の収集方法としては、検索対象となる情報を保持している情報資源に定期的にアクセスし、必要な情報を取得するという方法が一般的である。しかし、現在のように情報資源としての World-Wide Web サーバが氾濫する状況において、インターネットのような大規模なコンピュータ・ネットワーク上の全情報を単一ホストで管理するには限界がある。

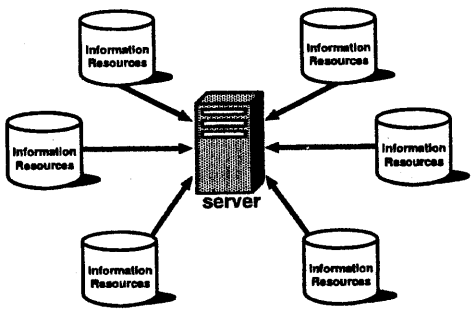

図 2.1: 一般的な検索システム

本研究では、これらの問題を解決する方法として分

散型自己情報管理方式とエージェントによるマイグレーション検索方式による情報共有システム (ATRAS: Agent-based Total Resource Access System)(図2.2)を提案する。本研究の究極的目標は、誰もが自由にかつ情報の存在する場所を気にすること無く、インターネット上の情報を利用することができる (これを情報共有と呼ぶ) ような環境を実現することである。

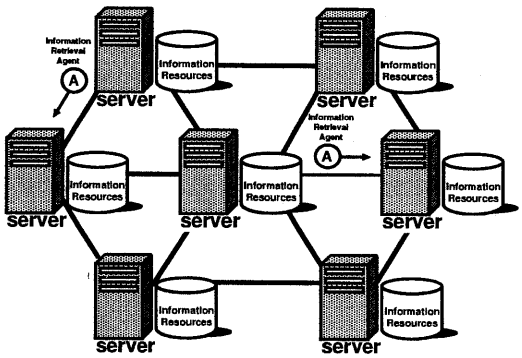

図 2.2: 本研究で提案する検索システム

2.2 システムの構成

インターネット上における本システムは、図 2.3 に示すような基本モデルの集合によって構成される。図2.3に示されるモデルは標準的な構成であり、この他にもう 2 種類モデルが存在する。また、図 2.3に示されている 5 種類のエージェントは本システムの最小構成要素であり、それぞれが異なる機能を持ち全体として一つの統合されたシステムを作り上げる。これらのエージェントは、情報検索の際にユーザとのコミュニケーションを行うユーザ・インタフェース・エージェント (Communicator)、実際にインターネット上で情報資源を探索し、目的の情報を見つけ出してくる情報探索エージェント (InfoSeeker)、その情報探索エージェントを管理する情報探索エージェント管理エージェント (SeekersManager)、情報資源から情報カタログを作成して検索依頼を処理する情報管理エージェント (InfoManager)、情報探索エージェントのホストへの出入りを監視するネットワーク・ゲートウェイ・エージェント (GateKeeper) である。本システムの基本概念は、これらのエージェントが相互にコミュニケーションを行い最終的にユーザの要求する情報を見つけ出してくることである。

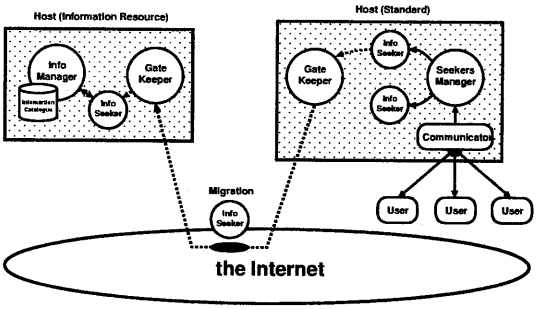

図 2.3: ATRAS 基本モデル

3 情報管理エージェント

情報管理エージェントは、本システムにおいて情報資源を提供するホストに常駐して、情報の管理、情報探索エージェントからの検索依頼の処理を行うエージェントである。本章では、情報管理エージェントによる情報の管理・検索方式について述べる。

3.1 情報資源

インターネット上で利用できる情報資源には様々な種類のものが存在する。従来からのテキスト・ベースでしかもコマンド・ラインから利用できる “whois” や “finger” サービスをはじめ、ドキュメント検索サービスの “WAIS(Wide Area Information Server)”、階層型メニュー方式による情報ブラウジング・サービスの “Gopher”、電子掲示版サービスの “NetNews”、ファイル転送サービスの “anonymous FTP” などが数多く存在する。その中で現在、非常に広く普及し利用されているのが “World-Wide Web” である。World-Wide Web は、サーバ構築の容易さとマルチメディア情報を扱えるということで、瞬く間に全世界のあらゆるユーザ層にまで広まった。

本システムでは、現在最も利用率が大きいであろうと思われる World-Wide Web と NetNews、anonymous FTP サービスに対象を限定して情報管理と情報検索を行う。

3.2 情報抽出

上記の情報資源に対して情報カタログを作成するために、各情報資源の持つ情報からその情報に関連する情報(その情報を導き出すという意味で以後キーワードと呼ぶ)を抽出しなければならない。すなわち、これらを行うには自然言語処理における意味理解や形態素解析などが必要となってくる。しかし日本語は英語と違い空白で区切られていないので自然言語処理、特に形態素解析に膨大な計算量を費す。また、本システムのように情報資源を提供する個々のホストが検索サーバとなるような構造に、情報管理のためだけに自然言語処理システムを実装するのは実用的ではない。

そこで辞書参照型のキーワード抽出法を考える。このアルゴリズムを図3.1に示す。

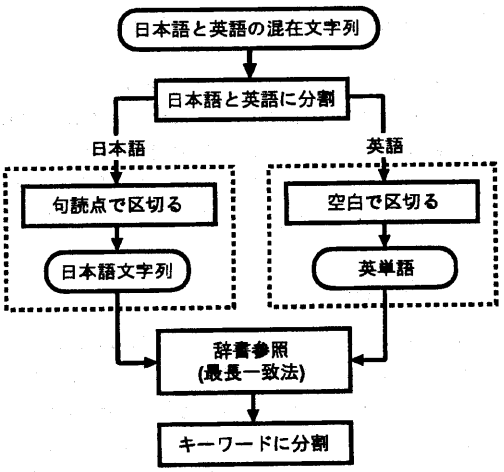

図 3.1: 辞書参照型キーワード抽出アルゴリズム

この方法の利点は、単純に辞書に登録されている単語だけをキーワードとして抽出するので、日本語形態素解析などのような膨大な計算量を必要としない。また、辞書はただのテキスト・ファイルなので単語の追加や削除が容易にできる。すなわち、特有の分野毎の固有名詞などを豊富に登録すればかなり精度の良い情報カタログの生成が期待される。さらに例外処理として片仮名文字列は全てキーワードとみなすようにしている。これは、一般に片仮名で表現される文字は外来語すなわち固有名詞であることが多いからである。

欠点としては、辞書に登録されていなければそれがどんなにその情報に関して特有のキーワードとなろうとも、情報カタログには登録されないという現象が生じることである。また、キーワード候補文字列を辞書から全検索するので、単純に辞書の大きさにキーワード候補文字列の照合時間が比例することになる。現時点では、文字列検索法として不一致文字法を用いたボイヤー・ムーア文字列探索法を利用している。これは最悪の場合でも辞書の大きさ N と文字列の長さ M の和 $N+M$ に上限が比例するだけである。

以上が、情報管理エージェントによる日本語文字列からの情報抽出法である。次に対象とする各情報資源毎の処理について説明する。

3.2.1 World-Wide Web

HTML(HyperText Markup Language) 文書におけるハイパーリンクは、テキスト・メディア以外の場合も存在する。その場合の情報を正しくカタログに登録するために、リンク元のアンカー文字列を情報カタログに登録する。ここでは、HTML 文書 X における情報 $I_{WWW}(X)$ を次のように定義する。

$$I_{WWW}(X) := T(X) + H(X) + U(Y) + A(Y) + B(X)$$

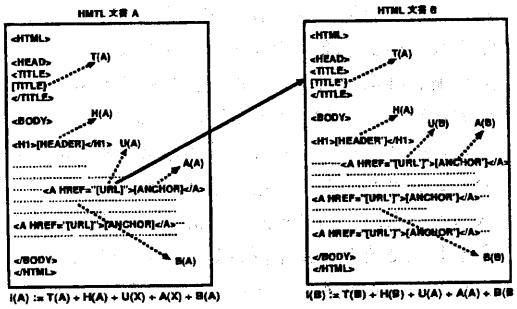

図 3.2: World-Wide Web からの情報抽出

ここで、$U(Y)$ は HTML 文書の URL(Uniform Resource Locators) であり、$A(Y)$ は HTML 文書 X にハイパーリンクしているリンク元のアンカー文字列である。また、$B(X)$ は HTML 文書 X からタ

グを取り除き、図3.1に示すアルゴリズムによって抽出されたキーワードの集合である。

3.2.2 NetNews

NetNews における情報の入手方法としては、普通は自分の興味のあるニュース・グループの記事を購読することで可能である。そこで NetNews の記事 X における情報 $I_{NEWS}(X)$ を次のように定義する。

$$I_{NEWS}(X) := S(X) + N(X) + M(X) + R(X) + B(X)$$

ここで、$B(X)$ はニュース記事の本文を対象に図3.1に示すアルゴリズムを用いて抽出されたキーワードの集合である。

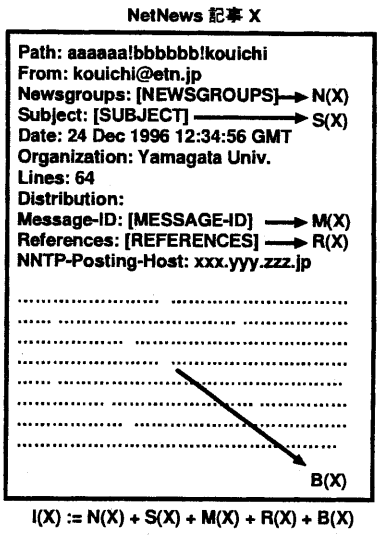

図 3.3: NetNews からの情報抽出

3.2.3 anonymous FTP

anonymous FTP における情報 (ファイル) の入手方法は、"archie" サーバと呼ばれる専用の検索サービスに接続して必要とするファイル名を入力することによって得られる。一般に、"archie" サーバで管理する情報カタログは図 3.4に示す形式を元に作成されるので、ファイル X における情報 $I_{FTP}(X)$ を次のように定義する。

$$I_{FTP}(X) := P_X(1) + P_X(2) + \ldots + P_X(n) + F(X)$$

ここで、$P_X(n)$ はパス中のディレクトリ名であり、$F(X)$ は対象とするファイル名である。

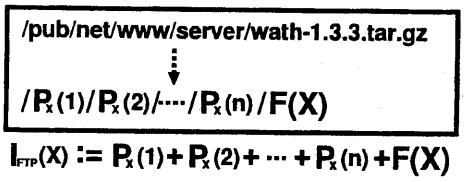

図 3.4: anonymous FTP からの情報抽出

3.3 情報の特徴付け

抽出した情報はこのままでは情報検索に利用できないので、情報検索に適した特徴付けが必要である。そこでここでは、ベクトル空間情報検索方式を導入することにする。この方式では、文書はベクトルとして表現される。ある文書ベクトル $\vec{D}$ を仮定すると、それぞれの要素 d_i はその文書に含まれる単語となる。このとき、それぞれの文書はベクトル $\vec{V}$ を持ち、その要素 v_i はその文書に含まれる単語 d_i の重みである。もし文書が単語 d_i を含んでいないならば、その重み v_i は 0 である。

単語の重みは TFIDF(Term Frequency × Inverse Document Frequency) 法を使用することにより決定される。最も簡単な場合は、文書 T 中の単語 d_i の重み v_i が次の式で与えられるときである。

$$v_i = tf(i) \cdot \log \frac{n}{df(i)}$$

ここで、$tf(i)$ は文書 T 中で単語 d_i が出現した回数 (用語頻度数)、$df(i)$ は d_i を含む文書の数 (文書頻度数)、n は収集した文書数である [2]。

3.4 情報管理

情報管理エージェントは、3.2節で抽出した情報について 3.3節で示した方法により情報の特徴付けを行う。その結果を情報カタログと呼ばれるインデクシング・ファイルに保存する。情報カタログ中では、全ての情報は情報資源の種類によらず統一した形式で管理される。以下にそのフォーマットを示す。

```
@KEYWORD   {
     WWW   { URL_1(w) ,  ··· ,  URL_n(w) }
    NEWS   { URL_1(w) ,  ··· ,  URL_n(w) }
     FTP   { URL_1(w) ,  ··· ,  URL_n(w) }
}
```

ここで、w は3.3節に示した方法で求められた単語(KEYWORD)の重みである。また、各情報資源に対応するURLは以下の形式を使用する。

- World-Wide Web:
  http://⟨host⟩:⟨port⟩/⟨path⟩?⟨searchpart⟩
- NetNews:
  news:⟨message-id⟩
- anonymous FTP:
  ftp//⟨host⟩:⟨port⟩/⟨path⟩

3.5 情報制限

必ずしも情報資源を提供するホストの全情報がいつも公開されているとは限らない。World-Wide Webサーバ上にページを作成したからといって、誰もが情報を不特定多数のユーザに公開したいと思っているわけではない。仲間内での情報交換の場として使用したりするために活用することもあるだろう。このような場合、検索システムに該当するページを登録させない方法が必要となってくる。現在、World-Wide Web上では"robots.txt"というファイルに検索システムに登録して欲しくないURLとサーチ・ロボット名を記述することで、この問題に対応している。情報管理エージェントによるWorld-Wide Webの情報管理でも、この規則を踏襲することにする。

よくNetNewsの記事として、無用なテスト記事やマルチポストされた広告記事などが投稿されることがある。情報管理エージェントは、これらの記事を情報としては何の価値もないものとみなす。そこでNetNewsの情報カタログ作成時にSubjectフィールドに日本語でも英語でもテストという意味の語が含まれている場合は、強制的に情報管理の対象から除外するようにする。

anonymous FTPでは、一般に"pub"というディレクトリで始まるパス以下にファイルが保存されているのが普通である。しかし、ホストによっては個人のディレクトリをanonymous FTPと同じレベルに作成している所もある。このような場合、個人のディレクトリを管理対象外にする必要がある。そこで情報管理エージェントでは、anonymous FTPの情報カタログ作成時にはWorld-Wide Webの"robots.txt"と同様な形式のファイルを利用することでこれに対処する。

3.6 情報検索

情報管理エージェントにおける情報検索は、3.4節で作成された情報カタログを用いて行われる。検索式は、以下に示す形式を取る。

```
     QUERY   :=   ⟨RESOURCES⟩ ⟨KEYWORDS⟩
⟨RESOURCE⟩   :=   [WWW | NEWS | FTP]
⟨KEYWORDS⟩   :=   ⟨meta⟩KEYWORD , ···
    ⟨meta⟩   :=   = , * , + , !
```

ここで、⟨RESOURCES⟩は検索対象となる情報資源に対応する文字を":"で区切って設定する。⟨meta⟩は検索式におけるメタ記号であり、KEYWORDは任意の文字列である。メタ記号は順に完全一致、and結合、or結合、not結合を意味する。基本的に検索は、文字列の部分一致方式であるため完全一致を示すメタ記号が含まれる。検索アルゴリズムは、現時点ではスコアリング方式を使用する。

4 おわりに

本論文では、分散環境における情報共有システムATRASの情報管理エージェントにおける情報抽出、情報の特徴付け、情報管理、情報制限、情報検索について詳細に述べた。現在の課題は、情報抽出・情報検索の高速化と情報カタログの質の向上である。今後は、これらの問題を解決しながらシステムの安定化を図りつつ実装を続けていく。

参考文献

[1] Michael Wooldridge and Nicholas R. Jennings, *"Intelligent Agents: Theory and Practice"*, Knowledge Engineering Review, October 1994.

[2] Marko Balabanovic and Yoav Shoham, *"Learning Information Retrieval Agents: Experiments with Automated Web Browsing"*, Department of Computer Science, Stanford University, Stanford, CA 94305.