知能と複雑系　125－4
（2001. 7. 23）

Webページを対象とした情報部品化手法の検討

荒谷 寛和 †　藤田 茂 †　菅原 研次 ‡
E-mail: aratani@sf.cs.it-chiba.ac.jp
† 千葉工業大学 情報工学科
‡ 千葉工業大学　情報ネットワーク学科

あらまし WWWの急速な増加によって，自然言語で表現された情報が多数存在する状況となっている．この多数存在する情報源から，目的となる情報を検索するために，yahooに代表されるような分類された索引を利用する，あるいはgoogleのような全文検索を利用することが行われている．情報検索の目的は様々なものがあるが，一般に検索を要求する状況に強く依存し，単純なテキスト一致や，事前に分類されたカテゴリからの検索では状況検索の要求に十分に答えることが困難である．本稿では，text/htmlで表現されたWebページを情報検索の対象として，そのページを一つのエージェントとすることで，Webページを一つの能動的な情報部品として捉えて，情報検索をエージェント指向で行う手法について述べる．

キーワード　**WWW　情報部品　ADIPSフレームワーク　協調プロトコル**

A Study on Making Agent Oriented Information Components from Web Pages

Hirokazu Aratani† Shigeru Fujita† Kenji Sugawara ‡
E-mail: aratani@sf.cs.it-chiba.ac.jp
† Department of Computer Science, Chiba Institute of Technology
‡ Department of Network Science, Chiba Institute of Technology

Abstract

There are much information on the WWW. Each information is written by natural language. The searcher who want to retrieve information from WWW use index page such as yahoo or search engine such as google. The aim of information search is depended on each situation, hence information retrieval as mechanically is not satisfied for the each requirement. In this paper, we described how to retrieve information from web page which is formated as text/html. Our approach is based on agent oriented information component framework. A web page will be change to an agent on our framework.

key words　**WWW　Information Component　ADIPS Framework　Cooperation Protocol**

1 はじめに

WWW の急速な増加によって，自然言語で表現された情報が多数存在する状況となっている．この多数存在する情報源から，目的となる情報を検索するために，yahoo に代表されるような分類された索引を利用する，あるいは google のような全文検索を利用することが行われている．情報検索の目的は様々なものがあるが，一般に検索を要求する状況に強く依存し，単純なテキスト一致や，事前に分類されたカテゴリからの検索では状況検索の要求に十分に答えることが困難である．一方，Semantic Web[1] のように記述された Web ページとメタデータを用いて，領域依存の知識表現を獲得/利用する研究が行われている．

雑多な情報が氾濫する Web ページを一意な情報空間と捉えて，全文検索や他サイトからの披リンク数をもとに得点付けし，1～3 単語からなる検索用キーワードから，有意な情報を万人向けに提供することは困難である．そこで，情報検索を行う作業者の嗜好や，情報検索の利用履歴を利用することが考えられる．しかしながら十分な数の利用履歴を蓄積するためには，作業者が多数の繰り返し利用を行い，検索結果に対して評価を入力する必要があり，作業者への負担が大きく，利用開始直後には十分な性能をシステムが示せないという問題がある．

Web ページに検索目的の情報が含まれているか，含まれていないかを判断するための基準として，(1) 単語が多く含まれている，(2) 他サイトからの披リンク数が多い，の 2 つを単純に採用することを考えると，(a) 常識的な情報が多く検索結果に出現する，(b) リンクページやニュースページのように関係の無い単語が単一の text/html で表現された Web ページに検索結果に出現する，(c) 掲示板に代表されるような話言葉で表現されており，多くの発言を読まないと主旨を理解できない Web ページ，が検索結果として出現することになる．

Web ページの構造や利用状況の分析からは，'ハブ &スポーク' という他の有用な情報へのリンクを持つ多くの作業者に利用されるサイトが，互いに繋がっているというモデルが存在することが示されている．これは index 型あるいは robot 型の検索サイトの実情を反映し，多くの常識的な情報を検索するためには有用であるが，ある特定の情報を欲している検索要求者に対して常に有効な手法ではない．

Web ページに記載された情報が特定のキーワードによって，常に一意に順序付けられるのではなく，ある仮定された情報を指示するページが検索対象となっている WWW 上で肯定されるか，否定されるかを Web ページの記載から推定することで，Web 上で多数の合意と支持を (まだ) 受けていない情報を検索することを実現する．

このために本稿では，text/html で表現された Web ページを情報検索の対象とし，Web ページを一つのエージェントと捉え，エージェント間の協調通信をもとにエージェント指向の情報検索を行う手法について述べる．

2 情報部品とその利用

2.1 Web ページの情報部品化

Web ページは妥当な HTML で表現されていることを仮定して，リンク情報を取得する．また，それぞれの Web ページの文章を形態素解析し，含まれる単語へのリンクとして単語の出現情報を取得する．

Web ページを情報部品として利用することは，(1)Web サイトが必ずしも単一のホストマシンにあることを想定できない，(2)Web サイトを構成する複数 text/html ファイル間に必ずしも強い相関があるわけでない，(3) 最終的に情報を求めている作業者に提示される情報が複数の text/html に渡って存在する場合でも情報部品の連結によって，情報を復元することが可能である，ということによる．

本稿で情報部品としての Web ページに要求する機能は，(a) 提示されたキーワードに関連する情報をもつか判定する，(b) 関連する情報を持つ他の Web ページを示す，(c) キーワードの提示にたいして真偽の判定を行う，(d) 他の Web ページの行った真偽の判定に対して支持/不支持の判定を行うの 4 点である．この機能を使って，情報部品の集合全体を用いて，提示されたキーワードへ自分自身/他者を推薦する，キーワードに情報を付加して真偽の判定を行う，他者の判断の信頼度を提示する．

これら 4 つの機能によって，情報部品の集合は作業者からの検索要求に対して，URL を推薦し，その URL に対する肯定，否定，支持，不支持を示す．次に情報部品の利用モデルについて述べる．

2.2 情報部品利用モデル

図 1 に，情報部品を使った情報検索システムの概念モデルを示す．

- text/html: Web ページを構成する要素として本システムで処理対象としている情報

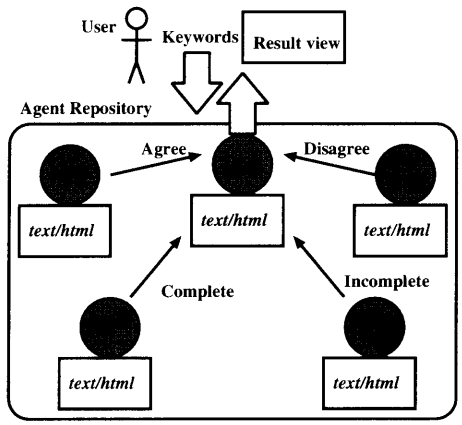

図 1: 情報部品利用システムの概念モデル

- Agent Repository: 情報部品の動作環境
- agent: ある text/html に対応する自律的オブジェクト
- User: 情報検索を行う作業者
- Keywords: 情報検索のキーワードとして，User から入力される単語列
- Result view: ある情報検索の結果として，Agent Repository が示す agent 間の協調結果としての情報一覧提示
- Agree, Disagree, Complete, Incomplete: 肯定，否定，情報として受け入れる，情報として受け入れられない，それぞれを表明する agent 間リレーション

本システムの動作順を以下に示す．

1. 作業者が検索用の単語列を入力する
2. リポジトリ内部のエージェントが単語列に対して検索結果の候補として名乗りをあげる
3. 名乗りを上げたエージェントの他に，他のエージェントに推薦されたエージェントが名乗りをあげる
4. 名乗りをあげたエージェント，それぞれに対してリポジトリ内部のエージェントが肯定，否定，受け入れ，受け入れ拒否のいずれかを表明する
5. 検索結果の表示として，名乗りを上げたエージェントと表明された意見をマージする

図 2 に，User に提示される情報一覧の表示モデルを示す．

Opinion は，検索用の単語列に対して直接関係すると Agent Repository の内部で名乗りを上げたエージェントへのリンクを示す．**Agree** は，そのエージェントに対して肯定の表明をしているエージェントへ，

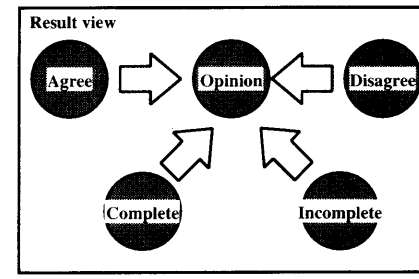

図 2: 情報検索結果の表示モデル

Disagree は，そのエージェントに対して否定の表明をしているエージェントへの，**Complete** は，そのエージェントの表明を受け入れられるエージェントへの，**Incomplete** は，そのエージェントの表明を受け入れることのできないエージェントへの，それぞれリンクを示す．

作業者に対して提示される URL は，Opinion が示す URL である．Opinion が複数存在する場合に，どのエージェント (Opinion) を作業者に提示するかは，Result view 作成を行う際のパラメータに依存する．Agree, Complete が多く，Disagree, Incomplete が少ない Opinion から順に提示するのがもっとも単純なパラメータ決定である．

2.3 情報部品の構成要素

WWW を構成する要素である text/html で表現された Web ページから，本稿で述べる情報部品として取り出す一次情報は，他の text/html へのリンク情報，単語の出現情報である．これに対して，情報検索用キーワードが与えられたときに情報部品自体が反応できたという情報と，そのときの Agree, Disagree, Complete, Incomplete を表明したエージェントの関係が二次情報として利用される．

情報部品は，一次情報としての部分と，二次情報としての部分から構成され，特に二次情報部分は，情報検索用のキーワードが与えられた毎に生成され，情報部品の利用毎に情報の更新を行う必要がある．本稿では，Web ページの作成者によって記述される text/html を一次情報本体，検索システムの動作によって蓄積される二次情報を情報システムを構成するエージェントによって管理し，個別の一次情報と一体化して，情報部品とする．以下にその構造を示す．

text/html はいわゆる html 形式で記述されたファイルに相当するテキスト情報であり，link は HTML

< 情報部品 > := < 一次情報，エージェント >
< 一次情報 > := <text/html, link>
< エージェント > := <agent, knowledge>

図 3: 情報部品

から抽出される URL 情報である．agent は 3 章で述べるエージェントであり，knowledge はエージェント動作のためのプログラムとその動作履歴である．

3 エージェントアーキテクチャ

情報部品化に用いられるエージェントは，我々が研究開発している ADIPS フレームワーク [2] で用いている構造を用いる．情報部品を構成するエージェントの内部構造を図 4 に示す．

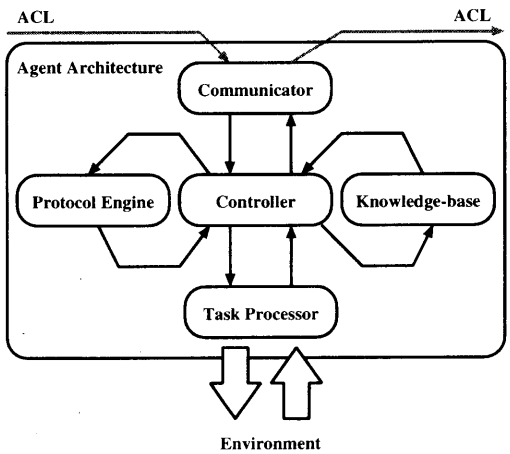

図 4: エージェントアーキテクチャ

- Controller
  エージェントを構成する要素をエージェント内部通信によって制御する機構．エージェントの動作を決定する．Controller を変更することで，エージェント間プロトコル，エージェントの知識を変更することなくエージェントの振舞いを変更する．
- Communicator
  エージェント間通信を実現する機構．エージェント外部から到着するメッセージをバッファリング，他のエージェントに対するメッセージを作成，適切なエージェントへの送信を行う[1]．
- Protocol Engine
  エージェント間プロトコルの処理機構．AgenTalk[4] に基づき，プロトコル記述とプロトコル処理部を分離する．エージェント間プロトコル記述をエージェントフレームワーク構築後に追加すること，エージェントの動作中での協調プロトコルの切替えを実現する．
- Knowledge-base
  個別のエージェントの知識処理機構．Protocol Engine と同様の設計思想により，推論エンジン部と知識記述部を分離している．
- Task Processor
  エージェントが他のエージェントでない計算機プロセスやファイル等の計算機環境を利用するためのインタフェース機構．いわゆる計算機プロセスをラッピングし，“エージェント化” するための API として機能する．

ADIPS フレームワークは，リポジトリに格納されたプログラム部品とそれに 1 対 1 で対応付けられたエージェント，およびプログラム部品を用いて分散システムを構成するエージェントから，作業者要求に基づいて分散システムを構成すること，およびその分散システムが状況変化に自律的に対応することを特長としている．今回はこのエージェントの構造を用いて，分散システムに代わって，情報検索結果を示すことで，情報検索システムとしての機能を実現している．すなわち，2.3 節での，情報部品は，ADIPS フレームワークにおける，プログラム部品に相当する部分が text/html, link の一次情報に相当し，作業者要求に基づき分散システムを構成，状況変化に適応するための知識記述が knowledge に相当する．

また，情報検索に対してエージェントが対応するための通信プロトコル，他のエージェントに対して肯定，否定，受け入れ，受け入れ拒否を表明するための通信プロトコルは，ADIPS フレームワークで定義している，Agent Communication Language, ACL から，基本プロトコルセット [3] を一部変更して用いる．情報検索に利用する ACL のパフォーマティブを図 5 に示す．

パフォーマティブ	意図	本稿での意図
information	情報通知	Opinion
agree	受諾	Agree
disagree	拒否	Disagree
complete	受信	Complete
incomplete	不完全	Incomplete

図 5: メッセージパフォーマティブ

[1] Communicator はさらにエージェントの動作環境上の通信支援機構を介して通信を実現する．本稿の内容とは直接関係しないため，割愛した．

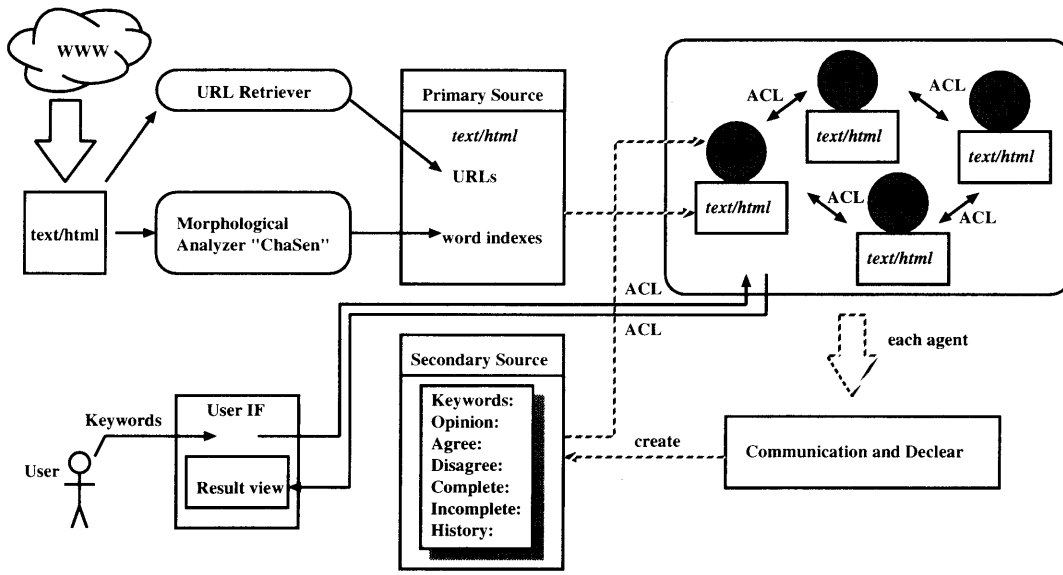

図 6: エージェント指向情報検索システム

4 エージェント指向情報検索システムの設計

text/html 情報の取得 robot 型検索サイトと同様に，不特定多数の Web サイトを巡回し，可能な限りの Web サイトをローカルな作業環境に保存する手法と，そのように構築されている既存の検索サイトを利用して，原 Web サイトの URL を特定する方法が考えられる．現時点では，ローカルな資源に制約があるので，後者を採用する．

また，原 URL からリンクを辿ることで，より有用な text/html を取得できることも多いことから，システムの機能として，エージェントが自律的に作業者から与えられたキーワードを補完する語を補い，さらに検索を行う場合がある．

一次情報の作成 作業者から入力されたキーワードに対して，Web ページを検索し，その原 URL をもとに検索対象となる text/html を得る．その text/html に含まれているリンク情報と，text/html の多くの情報であるテキストを "茶筅"[5] により形態素解析し，検索対象語のインデックスを作成する．

この text/html に含まれる語のインデックス情報とその語をキーとする URL へのリンク情報が一次情報として，エージェントから利用される情報源となる．

エージェントの動作と二次情報の作成 作業者から入力されたキーワードは，エージェントの集合に対して，ACL の形式で "関連する情報を持つか" という形式でブロードキャストメッセージで問い合わせが行われる．この問い合わせに対して，対応する text/html の語に対応する語句が存在するエージェントが次に，他のエージェントに対して，text/html の内容に対応する検索語のインデックスを示す．これは，エージェントが **Opinion** を表明することに相当する．

示された text/html に対して，エージェントは **Agree**, **Disagree**, **Complete**, **Incomplete** のいずれかを表明する．このとき，Agree を表明する場合は，(1) 同じように Opinion を表明している，もしくは，Opinion を表明しているエージェントのもつインデックスに対して，自分自身が Opinion を表明することのいずれかである．一方，Disagree を表明する場合は，Opinion を表明しているエージェントのもつインデックスに対して，自分自身のインデックスが否定するときである．Agree, Disagree が肯定/否定と内容に踏み込んだ判断をしているのに対して，Opinion を表明したエージェントに対して，その内容を肯定も否定もしないが，インデックスがキーワードに対して妥当であると判断する場合には，Complte を表明する．一方，Opinion を表明したキーワードとそのエージェントのインデックスに対して不適切であると判断した場合には，Incomplete を表明する．

エージェントの戦略 一回のキーワードの入力に対して全てのエージェントが 5 つの表明のなかのいずれかを必ずひとつを選択する．この動作履歴も 2 次情報として蓄積される．次回以降の動作では，エー

ジェントの知識記述 (Knowledge-base) により，あるキーワードに対して表明を行う際に，他のエージェントから多くの Agree と Complete を受け，できるだけ Disagree と Incomplete を受けないように，それまでの動作履歴を参照して Opinion を行うか，否かを決定する．

初期値としてエージェントは，同一の知識記述が行われている．しかし，動作履歴に基づいてキーワードに対する表明を変化させることで，より多くの Agree と Complete を受けるようにするために，同一の単語に対して Opinion を毎回表明するか，否かは，徐々に異なるようになる．

エージェントアーキテクチャ このように動作履歴に基づいて，エージェント自体の行動を変化させるためには，エージェント間通信プロトコル，知識記述に対して自らの動作パラメータを変更する機能が必要であり，このため ADIPS フレームワークにおけるエージェントアーキテクチャを用いる．3 節で述べた Protcol Engine および，Knowledge-base がプロトコル記述部と知識記述部を分離し，動作中の動的な変更を可能にするエージェント全体の Controller があることで，エージェントは動作中に得る履歴をもとに，それまでの ACL への反応とは異なる ACL を出力する．

この結果，他のエージェントから自分自身の Opinion に対しての表明をもとに，利用する協調プロトコルを切替える (他のエージェントへの表明の切替え)，推論に利用する知識記述のモジュールを変更する (Opinion，他のエージェントへの表明の切替え) ことで，初期値とは異なる振舞いを見せることを実現する．

view の生成 全てのエージェントがあるキーワードに対して，表明を終えた後に，作業者に対して検索結果の一覧 (view) を表示する必要がある．一般に多くの検索サイトでは，ある基準にしたがって，上位数十件毎に順次一覧を作成することで，利用者に対して検索結果を示している．本システムでは，Opinion を表明したエージェントが直接の検索結果に相当する．このエージェントの順位付けは，Agree, Disagree, Complete, Incomplete の組合せによって変更可能であり，作業者からの支持に基づいて Opinion を表明したエージェントを並び替えて示すことになる．

作業者の評価反映 作業者に対して示された view が，その作業者が求めてる情報本体に近いものであるか，否かを逐次入力することで，個別の作業者毎の利用者モデルを構築することが考えられる．作業者が入力しなかったキーワードや，最終的に利用する Web サイトが特定できるような場合 (ex., ニュース専門サイトなど) には，該当する Web ページに相当するエージェントが，Opinion として表明するように戦略を変更する．

5 おわりに

本稿では，text/html で表現された Web ページを情報検索の対象とし，Web ページを一つのエージェントと捉え，エージェント間の協調通信をもとにエージェント指向の情報検索を行う手法について述べた[2]．

Web ページのコンテンツに対して，単純にリンク数が多い場合にその Web ページの評価を上げるのではなく，Web ページ自体のコンテンツと評価する Web ページのコンテンツを比較することで，そのコンテンツを肯定，否定，支持 (受入れ)，不支持 (受入れ不能) と評価し順序付けるシステムとして，検索システムを設計した．また，キーワードに対する検索結果として自分自身が妥当であると主張したときの，他の Web ページからのコンテンツの評価を反映することで，他のコンテンツの内容との関連をコンテンツ自体が獲得し，次の検索に反映させる仕組みを持たせた．

本システムは現在設計に基づいて Java で実装中であり，引続き評価実験を予定してる．

参考文献

[1] Dieter Fensel, Mark A. Musen **ed.**, "The Semantic Web: A Brain for Humankaind", IEEE Intelligent Systems, March/April, 2001

[2] 藤田，他，"分散処理システムのエージェント指向アーキテクチャ"，情報処理学会誌，Vol.37, No.5, pp.840–852，1996 年

[3] 藤田，"協調プロトコルの混在を目的とした ADIPS フレームワークにおけるエージェントアーキテクチャの設計"，信学技報，AI2001-41,2001 年

[4] 桑原，他，"Agentalk: マルチエージェントシステムにおける協調プロトコル記述"，電子情報通信学会論文誌 B-I，Vol.J79-B-I, No.5, pp.346 – 354,1996

[5] "茶筅"，http://chasen.aist-nara.ac.jp/

[2] 本研究の一部は，科研費:13780340．および，学振未来開拓事業:動的ネットワーキングによる．