動的情報メディエータのための知的情報収集手法

北村　泰彦[†]　　野田　智哉[††]　　辰巳　昭治[†]

An Intelligent Information Gathering Techinique for Dynamic Information Mediator

Yasuhiko KITAMURA[†], Tomoya NODA[††], and Shoji TATSUMI[†]

あらまし
今日，インターネットは社会に急速に普及し，われわれの生活に欠かすことのできない基盤の１つになりつつある．中でも，WWW は個人の情報発信，学術研究，企業活動や電子商取引など様々な目的で利用され，発信される情報量は急速に増えつづけている．これらのイン ターネット情報源は単独で使用されることが一般的であったが，複数の情報源を組み合わせて利用することができればその価値をより高めることができるので，インターネット上に分散した情報を統合する情報メディエータが注目されている．本稿では，統合すべき情報源が頻繁かつ非同期に更新されるという前提のもとで，利用者のクエリに対して，許容時間内に情報源を適切にアクセスして解を導き出すという情報収集技法を提案する．本方式では，キャッシュ内データの信頼性と導出すべき解の質を考慮して，適切な情報アクセスと解の提示ができる．本方式の評価としては，実世界の WWW サイトからの情報を用いた航空情報メディエータを構築し，従来の FIFO 型キャッシュ手法と比較して，より効率よく情報アクセスを行っていることを明らかにした．

キーワード　インターネット，WWW，情報メディエータ，キャッシュ，情報統合

1. はじめに

今日インターネットは社会に急速に浸透し，我々の生活に欠かすことのできないインフラストラクチャの１つになりつつある．中でも WWW は，電子商取引をはじめとした企業活動，学術研究，コミュニティ形成など様々な目的で利用されている．これらの WWW 情報源はそれぞれ単独で利用するのが一般的であるが，これらを組み合わせて利用することができれば，その利用価値が高められる．

このための手段として，リンク集やポータルサイトでは，関連した WWW 情報源をハイパーリンクで連結させることで関連付けを行っている．また，サーチエンジンは，利用者が入力したキーワードを含む WWW ページの一覧を動的に作成して表示する．しかしながらこれらの手段は，関連するページの集合を提供するだけであり，ページの収集，必要な情報の抽出，抽出された情報の統合といったことは自動的には行われない．そこで，複数の WWW 情報源から発信される情報を組み合わせて，問題解決に利用する情報統合技術が重要になる．

WWW 情報統合の例としては，航空機の空席照会システムが考えられる．日本の主要な航空会社は，運行スケジュールと空席状況を検索できるサービスを WWW で提供している．このサービスでは搭乗日，出発地，目的地などを入力すると，該当する便の便名，運行スケジュール，空席状況などが結果として表示される．しかしながら複数の航空会社が同一経路を運行している場合には，それぞれの航空会社のサイトにアクセスする必要がある．また乗り継ぎ情報が必要な場合も繰り返しサイトにアクセスする必要が生じる．そこで，複数の航空会社の提供するサービスの結果を統合して出力する空席照会システムを実現できれば，利用者はこのようなわずらわしさから解放される．

しかしながら WWW 情報統合を実現するには以下

† 大阪市立大学工学部，大阪市
Faculty of Engineering, Osaka City University, Sugimoto 3-3-138, Sumiyoshi-ku, Osaka, 558-8585 Japan
†† NTT コムウェア，札幌市
NTT Comware, Ohdohrinishi 7-3, Chuo-ku, Sapporo, 060-0042 Japan

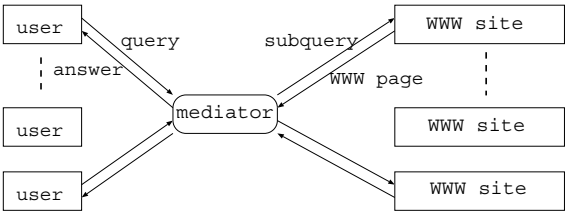

図 1　メディエータによる情報統合
図 2　Infromation integration through a mediator.

の問題点がある．

(1) 自律分散した WWW 情報源： WWW 情報源はインターネット上に分散して存在し，それぞれが独立して管理，運営されている．そのため分散している情報源から，関連している情報を収集しなければならない．

(2) アクセスコスト： WWW 情報源へのアクセスはインターネットを介するので時間がかかる．また有料の情報源へのアクセスは費用がかかる．特に，WWW 情報統合では多くのの WWW ページを収集する場合もあり，アクセスにかかる時間やコストはさらに大きくなる．

(3) 動的な WWW 情報源： WWW 情報源の中には頻繁にに情報の更新がおこなわれるものがある．また，更新の頻度はそれぞれの情報源で異なり，非同期に行われるので，更新されたかどうかは実際にアクセスしてみないと分からない．

(1) の問題に対処するために，本稿では図 1 に示すように分散している情報を統合する情報メディエータ [15] を採用している．利用者からクエリ (query) を受け取ると，メディエータは複数の WWW 情報源にアクセスして必要な情報を抽出する．そして得られた情報を用いて，利用者のクエリに対する解 (answer) を求め，利用者に返す．次に (2) に対してはキャッシュ用いることで対処する．一度得た情報を収集してキャッシュしておくと，WWW 情報源にアクセスする回数が減るので，応答時間を短縮することができる．

しかし，(3) に示されるように，WWW 情報源の発信する情報は頻繁に更新される可能性がある．そのため，キャッシュした情報が古くなってしまうと正しい解を提示できなくなってしまう．一方，キャッシュを利用しなければ，大量の情報収集を必要するときには，解の提示に長時間を要してしまう．そのため，限られた時間内でいかに適切に情報収集を行うかが重要な課題となる．

そこで，本稿では，過去に得た情報をメディエータ内にキャッシュしつつ，利用者のクエリが来た後に必要な情報の再検索を効率良く行うことで，適切な解の提示を行う手法を提案する．そのために，キャッシュした情報の信頼性と質を考慮して再検索すべき情報を決定し，ネットワークへのアクセス回数が限定された状況下においてもできるだけ適切な解を提示する情報収集アルゴリズムを提案する．

2 章では我々の提案するアルゴリズムについて説明し，3 章では実験によって我々のアルゴリズムの有効性を示す．4 章で関連研究について述べ，5 章でまとめと今後の課題とする．

2. 知的情報収集アルゴリズム

本稿では WWW 情報統合を実現するために図 1 に示したメディエータを採用する．メディエータは利用者からのクエリに対して，利用者の代わりに必要であれば複数の WWW サイトにアクセスし，その情報を抽出，統合することで得られた解を利用者に提示する．メディエータのアクセスコストを軽減するために，一度得たデータはメディエータ内にキャッシュされる．しかしながら，情報源のデータが頻繁に更新される場合には，一度キャッシュしたデータであっても，再検索する必要が生じる．一方，利用者に対してはある許容時間以内に何らかの解を返す必要があり，再検索はこの許容時間内に行う必要がある．すなわち，どの情報源に対して再アクセスするかが重要な課題となる．

本節では，キャッシュするデータの信頼度と質を考慮して再検索するデータを決定する知的情報収集アルゴリズムについて述べる．このアルゴリズムでは，許容時間内でデータの再検索を繰り返し行い，得られたデータの内容によって次に再検索するデータを動的に決定する．

2.1 事　　実

メディエータは WWW ページを収拾し，そこから必要なデータを抽出する．ここでは抽出される原子的な情報を事実 (fact) と呼ぶ．単一の WWW ページから複数の事実が抽出されることもある．例えば図 3 で示されるような飛行経路を扱う空席照会システムでは，事実として図 4 のような運行スケジュールと空席状況が抽出される．メディエータは事実を抽出すると，それをキャッシュしておく．

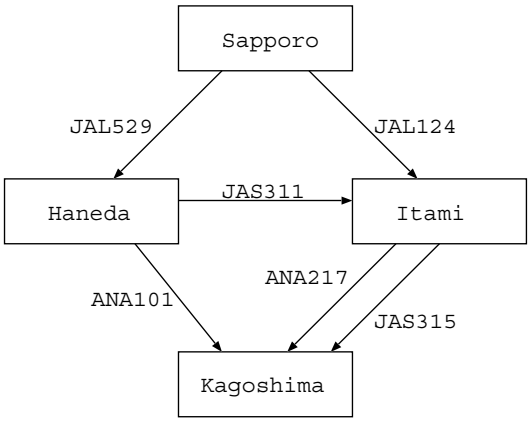

図 3 飛行経路
Fig. 3 Flight connection.

Fact	Reliability
flight(JAL124,Sapporo,Itami,0800,0900)	1
flight(JAL529,Sapporo,Haneda,0700,0800)	1
flight(ANA217,Itami,Kagoshima,1000,1200)	1
flight(JAS315,Itami,Kagoshima,1300,1500)	1
flight(JAS311,Haneda,Itami,0830,0930)	1
flight(ANA101,Narita,Kagoshima,0930,1100)	1
availability(JAL124,1999/12/16,Yes)	0.90
availability(ANA217,1999/12/16,Yes)	0.60
availability(JAS315,1999/12/16,No)	0.30
availability(JAL529,1999/12/16,Yes)	0.70
availability(JAS311,1999/12/16,Yes)	1.00
availability(ANA101,1999/12/16,No)	0.60

図 4 空席照会システムにおける事実
Fig. 4 Facts for flight information service.

2.2 事実の信頼度

情報源の発信する情報が頻繁に更新される場合には，時間がたつとともにメディエータがキャッシュした事実と，もともとの情報源の情報に食い違いが生じる可能性が大きくなる．そこで，メディエータがキャッシュした事実に対し，その事実がどの程度信頼できるかを表すために，0 から 1 の値をとる信頼度 $r(f)$ を定義する．信頼度は情報源から事実を抽出してキャッシュした直後が最も値が高く，時間が経つにつれてその事実の信頼度は低下する．そのため $r(f)$ は，事実 f をキャッシュしてからの経過時間に対して，1 から 0 に収束する単調減少関数になる．

それぞれの事実ごとに更新頻度は異なるため，信頼度関数のふるまいはそれぞれの事実に依存する．例えば，航空機の空席状況，株価や為替相場のように非常に頻繁に更新される事実であれば，その信頼度はすぐに低下する．一方，航空機の運行スケジュールのようにまれにしか更新されない事実も存在し，このような事実の信頼度は長時間一定である．

信頼度関数をここでは

$$r(f) = \frac{1}{1+wt}$$

として近似する．ここで，t はキャッシュしてからの経過時間，w は事実のタイプに依存する重みである．w を大きくすると信頼度は早く減少し，小さくするとゆっくりと減少する．

本稿では，更新される事実を動的な事実 (dynamic fact)，更新されない事実を静的な事実 (static fact) と呼んでいる．実世界において絶対に更新されない事実というのはわずかだと考えられるが，扱う問題に応じて，更新されないとみなしても支障のない事実を静的な事実とする．例えば空席照会システムであれば，非常に頻繁に更新される空席状況が動的な事実になる．一方，ほとんど変更されない運行スケジュールは静的な事実とみなせる．静的な事実は一度キャッシュすると，再検索する必要はなく，再検索の対象は動的な事実にしばられる．

2.3 解

利用者からクエリを受け取ると，メディエータはキャッシュしている事実を用いて，利用者のクエリに対する解 (answer) を導出する．具体的な導出方法は問題によって異なるが，例えば以下のように Prolog 的なルール表現できる．

$(R1)$ query($A,$B) :− $fact($A,$B).

$(R2)$ query($A,$B) :− $fact($A,$C), query($C,$B).

ここで$fact($A,$B) はキャッシュされている事実であり，$A と$B は変数である．(R1),(R2) を繰り返し適用することでクエリを満たす事実の集合が得られる．一般に解は複数個存在する．

空席照会システムであれば，解を導出する手続きは，以下のように表現できる．

$(R3)$ query($P1,$P2,$Date,$T1,$T2) :−
$flight($Number,$P1,$P2,$Dep,$Arr),
$availability($Number,$Date,$Status),
$Dep >= $T1,
$T2 >= $Arr.

$(R4)$ query($P1,$P2,$Date,$T1,$T2) :−
$flight($Number,$P1,$P3,$Dep,$Arr),
$availability($Number,$Date,$Status),
$Dep >= $T1,
$T2 >= $Arr,
query($P3,$P2,$Date,$Arr,$T2).

(R3) は日付 Date に時刻 T1 以降に空港 P1 を出発し，空港 P2 に時刻 T2 までに到着する直行便を表す．(R4) は空港 P3 を経由するルートである．例えば，札幌から鹿児島にいく旅程 query(Sapporo,Kagoshima, 1999/12/16,0600,1600) を利用者が要求した場合には，図 4 で示されるキャッシュされたデータから，図 5 で示される解が導かれる．

2.4 解 の 質

解の質 (quality) は，その解がどの程度利用者のクエリを満たしているかを表す．質は，再検索する事実や，利用者に提示する解を決定するために使用される．解の質は解を構成する事実から計算され，動的な質 (dynamic quality) と静的な質 (static quality) に分けられる．動的な質は動的な事実に関する質であり，静的な質は静的な事実に関する質である．

それぞれの質の具体的な評価方法は，扱う問題によって異なる．空席照会システムでは，運行スケジュールと空席状況が事実になるので，これらを用いて質を評価する．利用者の希望する経路の飛行時間が短く，かつ用いるすべての便に空席があれば，良い解といえ，そのような解の質は高くすればよい．この例では静的な質は飛行スケジュールに関するものなので，解 A の静的な質は以下のように定義できる．

$$Q_S(A) = \begin{cases} 1 - \dfrac{t_a - t_d}{24} & 0 \le t_a - t_d \le 24, \\ 0 & \text{otherwise}, \end{cases}$$

ここで t_a は希望する到着時刻，t_d は最初のフライトの出発時刻である．すなわち 24 時間以内の旅行時間で，それが短いものほど質が高いといえる．それ以外の質は 0 と定義されている．

動的な質は空席に関するものなので，以下のように定義できる．

$$Q_D(A) = \begin{cases} 1 & \text{if 全てのフライトに空席がある}, \\ 0 & \text{otherwise}. \end{cases}$$

解全体の質は以下のように静的な質と動的な質の積を取ることにより定義している．

$$Q(A) = Q_S(A) \cdot Q_D(A).$$

例えば，図 5 より，札幌から鹿児島に行くルートは 5 種類存在し，到着時刻は 11:00，12:00，15:00 の 3 種類が存在する．もし利用者が 16:00 までに鹿児島に着きたいならば，A_1 の静的な質は

$$Q_S(A_1) = 1 - \frac{16 - 8}{24} = 0.67$$

となる．動的な質は JAL124 と ANA217 がともに空席があるので，1 である．A_2 の静的な質は A_1 と同じであるが，動的な質は JAS315 が満席であることから，0 となる．したがって解全体の質は $Q(A_1) = 0.67$，$Q(A_2) = 0$ となる．

2.5 解の信頼度

メディエータは，自身のキャッシュしている事実を用いて解を求めるが，解の導出の用いた事実の信頼度が低いと，その解の信頼度も低いと言える．そこで，解の信頼度 $R(A)$ を以下のように定義する．

$$R(A) = \min_{f \in A} r(f)$$

例えば，解 A_1 の信頼度は以下のように計算される．

$$R(A_1) = \min\{0.90, 0.60\} = 0.60.$$

2.6 アルゴリズム

解はメディエータのキャッシュする事実の集合を用いて導出するが，WWW 情報源の発信する情報は更新されている可能性があり，解の信頼度は低下する．解の信頼度を向上させるには WWW 情報源より再検索すればよいが，それは利用者の許容時間内で行う必要がある．したがって，再検索する事実の決定方法が重要になる．そこで解 A に対して，再検索する事実を決定するための再検索スコア $S(A)$，利用者に提示する順序を決定するための提示スコア $P(A)$ を定義し，これらの関数を用いて，再検索を行う事実や利用者に提示する解を決定する．図 6 にそのアルゴリズムを示す．

アルゴリズムについて説明する．最初に，利用者からクエリ (query) を受け取ると，キャッシュされた事実を用いて解を導出する (1 行目)．(注1)次に，できる

(注1)：ここでは全ての事実はあらかじめキャッシュされており，静的な事実と動的な事実の信頼度はそれぞれ 1 と 0 に設定されていると仮定している．

Answer	Facts
A_1	flight(JAL124,Sapporo,Itami,0800,0900), availability(JAL124,1999/12/16,Yes), flight(ANA217,Itami,Kagoshima,1000,1200), availability(ANA217,1999/12/16,Yes).
A_2	flight(JAL124,Sapporo,Itami,0800,0900), availability(JAL124,1999/12/16,Yes), flight(JAS315,Itami,Kagoshima,1300,1500), availability(JAS315,1999/12/16,No).
A_3	flight(JAS529,Sapporo,Haneda,0700,0800), availability(JAS529,1999/12/16,Yes), flight(JAS311,Haneda,Itami,0830,0930), availability(JAS311,1999/12/16,No), flight(ANA217,Itami,Kagoshima,1000,1200), availability(ANA217,1999/12/16,Yes).
A_4	flight(JAS529,Sapporo,Haneda,0700,0800), availability(JAS529,1999/12/16,Yes), flight(JAS311,Haneda,Itami,0830,0930), availability(JAS311,1999/12/16,No), flight(JAS315,Itami,Kagoshima,1300,1500), availability(JAS315,1999/12/16,No).
A_5	flight(JAS529,Sapporo,Haneda,0700,0800), availability(JAS529,1999/12/16,Yes), flight(ANA101,Narita,Kagoshima,0930,1100), availability(ANA101,1999/12/16,Yes).

図 5　query(Sapporo,Kagoshima,1999/12/16,0600,1600) に対する解の集合
Fig. 5　A complete set of answers to query(Sapporo,Kagoshima,1999/12/16,0600,1600).

```
1: construct answers {A_1, A_2, ···} to a user's query by using cached facts
2: repeat until query time expires {
3:     select A_i where S(A_i) = max{S(A_1), S(A_2), ···}
4:     select f_m where r(f_m) = min_{f_n ∈ A_i} r(f_n)
5:     reload a WWW page and update f_m }
6: sort answers by P(·)
7: return the best answer to the user
```

図 6　情報収集アルゴリズム
Fig. 6　Information gathering algorithm.

だけ正しいな解を提示するために，メディエータは予め指定された時間が経過するまで，事実の再検索を行う (2-5 行目)．繰り返しの中では，更新すべき事実を選択する (3 行目)．それは各解 A_i に対して，

$$S(A_i) = Q_S(A_i) \times (1 - R(A_i))$$

を計算する．この式により，静的な質が大きく，信頼度の低い解を選択することができる．そのような解は更新されれば，信頼度と質がともに高い解に至る可能性が高い．一方，静的な質の低い解は更新されたとしても，質は高くないので考慮する必要はない．ここでは質に関しては静的なもののみを考慮しているが，動的な質は検索するまでその値が未知であるからである．

次に選択された解の中から，最も信頼度の低い事実が選択され (4 行目)，最終的にその事実を更新するために WWW サイトから関係するページを再検索し，事実の更新を行う (5 行目)．

再検索の時間が過ぎると，メディエータは解を $P(A_i)$ によりランク付けする．$P(A_i)$ は

$$P(A_i) = Q(A_i) \times R(A_i)$$

として定義され，信頼度と質の高い解を反映するようにしている (6,7 行目)．

ここで表 1 を用いてどのように事実が更新されるかを示そう．最初に 5 つの候補の中から S 値が最も大きい解 A_2 が選択される．A_2 には現在，満席のフライト JAS315 が含まれている (図 5 参照) が，図 4 に示されるようにその信頼度 (Reliability) は低い (0.30) ので，その事実はすでに更新されているかもしれない．したがって， availability(JAS315,1999/12/16,No) が更新のために選択される．ここでこの事実が availability(JAS315,1999/12/16,Yes) に更新されているとしよう．このとき，その信頼度は 1 になり，この事実を含む解 A_2 と A_4 が更新される．

2 回目の更新の際には S 値が最大になる A_1 が選択され，これまでと同様に availability(ANA217,1999/12/16,Yes) が更新される．今度は availability(ANA217,1999/12/16,No) になったと仮定しよう．このとき A_1 と A_3 の Q 値が 0 になる．この時点で更新時間が過ぎたとすると，最大の P 値をもつ解 A_2 が選択され，利用者に提示される．

3. 評価実験

理想的なメディエータは最適な解を短い時間で返すことができると考えられる．そこで，ランダムに作成したクエリをメディエータに送り，再検索の回数のを変化に応じて，メディエータの提示する解と最適解が一致する確率 (成功率) がどのように変化するかを測定した

表 1　事実の更新に応じた各関数値の変化
Table 1　Changes of function values according to the number of updates

A_i	Q_s	No update				1st update				2nd update			
		Q	R	P	S	Q	R	P	S	Q	R	P	S
A_1	0.67	0.67	0.60	0.40	0.27	0.67	0.60	0.40	0.27	0.00	0.90	0.00	0.07
A_2	0.67	0.00	0.30	0.00	0.47	0.67	0.90	0.60	0.07	0.67	0.90	0.60	0.07
A_3	0.63	0.63	0.60	0.38	0.25	0.63	0.60	0.38	0.25	0.00	0.70	0.00	0.19
A_4	0.63	0.00	0.30	0.00	0.44	0.63	0.70	0.44	0.19	0.63	0.70	0.44	0.19
A_5	0.63	0.00	0.60	0.00	0.25	0.00	0.60	0.00	0.25	0.00	0.60	0.00	0.25

比較対象には，キャッシュした事実の質を考慮せずに，キャッシュの信頼度のみを再検索する事実の決定に用いる FIFO(First-In First-Out) 法(注2)を用いた．これは，最も古くキャッシュされた事実から再検索を行う方法である．これは提案アルゴリズムにおいて，再検索スコアを

$$S_{FIFO}(A) = 1 - R(A)$$

とした場合と等価である．

評価実験は実世界の航空会社の WWW サイトからの情報を統合する空席照会システムの上で行った．これらの WWW サイトは JAL (Japan Air Line)(注3), ANA (All Nippon Airways)(注4), JAS (Japan Air Systems)(注5)により提供されているものである．ここでは運行スケジュールを静的な事実，空席情報を動的な事実と見なしており，2 節で述べたように解や事実の質と信頼度を計算している．

評価実験では，1999 年 12 月 18 日に，札幌，羽田，伊丹，関西，福岡，那覇空港を離発着する JAL，JAS，ANA の便を対象とした．1999 年 12 月 8 日から，12 月 18 日まで，出発地と到着地をランダムに作成したクエリーを 1 時間間隔で送り，最適解を提示する確率 (成功率) を測定した．(注6)乗り継ぎの回数は 1 回に制限した．これは実世界でも実行可能なように問題の規模を小さくするためである．最適解を得ることのできる比率を図 7 に示す．横軸は更新の回数を表し，縦軸は最適解を得た比率を表している．

両手法ともに更新の回数が増えるにしたがって成功率は向上している．おおまかに述べると，一つのクエリーに対して 8 回のアクセスで提案手法 (IIGA) は最適解を得ることができる．一方 FIFO は約 18 回のアクセスが必要であることがわかる．これは FIFO が解の信頼度のみを考慮して，再検索する事実を選択しているのに対して，提案手法では解の信頼度と質をともに考慮していることによるものである．この評価実験では，実世界の環境でも評価可能なように問題の規模をかなり小さくしている．乗り継ぎの回数を 1 に制限しているので，解を構成する事実の数は高々4（静的な事実と動的な事実がそれぞれ 2）である．さらに，航空会社の提供する WWW サイトでは一つのクエリーに対して多数の事実を含むページを返すことが多く，これも全体的に WWW アクセスを減らす要因にもなっている．(注7)しかしながら問題の規模をさらに拡大するならば，提案手法の優位性はより顕著なものになると予想される．

4. 関連研究

情報統合の研究はデータベースや人工知能の分野において盛んに行われている．[4], [7], [10], [16] データベースからの代表的な事例として，米国 Stanford 大学の Tsimmis プロジェクトは WWW 情報源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 [2]．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エータ (mediator) [15], [17] により統合する点である．本研究においてもメディエータを情報統合の手段として採用している．また人工知能からのアプローチとしては，情報エージェントとファシリテータを ACL(Agent Communication Language) [3] 用いて統合する連邦アーキテクチャ [6] がある．

(注2)：他のよく知られたキャッシュ戦略としては LRU(Least Recently Used) があるが，これは静的な情報源を前提としており，本論文で対象となっている動的な情報源のためのキャッシュ更新アルゴリズムとしては適さないので，比較の対象とはしなかった．
(注3)：http://www.5971.jal.co.jp/cgi-bin/db2www/avail.d2w/report
(注4)：http://rps.ana.co.jp/drs/vacant1.cgi
(注5)：http://www.jas.co.jp/kusektop.htm
(注6)：全部で 264(=24×11) のクエリーを送ったが，WWW サイトの障害のために，そのうち 213 のクエリーに対してメディエータは何らかの解を導出した．
(注7)：例えば，ANA のサイトでは単一のクエリーに対する応答のページから 22 の事実を抽出することができた．

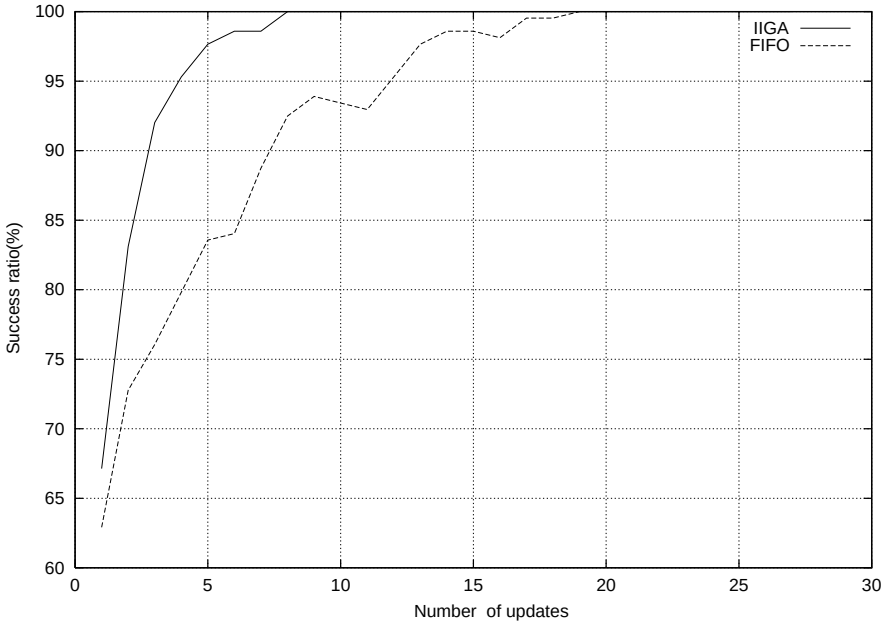

図 7 Experimental Result

従来の WWW 情報統合は主に WWW 情報源の構造化 [11]，WWW 情報源からの情報抽出 [8] や収集 [5], [13]，情報マッチメイキング [12] といった話題が中心に議論されてきた．しかしながらこれらの研究は文献 [1] を除いて，情報源が頻繁に更新されるような前提では議論されてこなかった．文献 [1] では分散メディエータシステムにおけるキャッシュと最適化の問題を扱っている．ここでは主に，通信オーバヘッド，情報源の応答時間，経費，障害などを考慮して，どの情報サイトにアクセスすべきか，またキャッシュすべきかについて議論している．それに対して本論文では，通信路やサイトの性能には特に着目せず，解の信頼性と質を考慮して，限られた時間の中でできるだけ良い解を提示するための情報収集アルゴリズムを議論している．view maintenance [18] の研究では複数の情報源からの情報を無矛盾に統合しようとしている．しかしながら分散している大量の情報を無矛盾に保つためには大量の通信が必要になると考えられ，また情報源が頻繁に更新される場合にはそれを行うことは不可能に近くなる．本論文の手法は限られた検索時間内でできるだけキャッシュの内容を無矛盾にするような半矛盾的 (semi-consistent) なアプローチを取っているといえる．

ここで提案されたアルゴリズムは常に何らかの解が得られ，時間をかければかけるほど良い解が得られるというエニータイムアルゴリズム [14] の性質も兼ね備えている．

5. ま と め

WWW 情報統合の問題点として情報源の更新の問題に対処する方法を提案した．メディエータは限られた時間内に適切に情報を収集し，適切な解を利用者に提示する必要がある．ここではキャッシュ内のデータの信頼度と解の質を考慮し，再検索すべき事実を選択する知的情報収集アルゴリズムを提案した．従来の FIFO では信頼度ののみしか考慮していなかったのに対して，この手法では解の質を考慮するために，FIFO よりも優れた性能を示すことを空席照会システムという現実問題に応用することで示した．ここでは具体的な応用事例として航空機の空席照会システムを取り上げたが，本手法は頻繁に更新されるような情報源の統合に有効であると考えられる．株価情報システムや電子オークションなどの他の領域への応用も検討してゆきたい．

謝辞

本研究の一部は，新エネルギー・産業技術総合開発機構 (NEDO) の委託事業「シニア支援システムの開発」によるものである．

文　献

[1] Adali, S., Candan, K.S., Papakonstantinou, Y., and Subrahmanian, V.S.: Query Caching and Optimization in Distributed Mediator Systems, SIGMOD-96 (1996) 137–148

[2] Chawathe, S., Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J. and Widom, J.: The TSIMMIS Project: Integration of Heterogeneous Information Sources. Proceedings of IPSJ Conference (1994) 7–18

[3] Finin, T., Labrou, Y., and Mayfield J.: KQML as an Agent Communication Language. Bradshow, J.M. (ed.): Software Agents. AAAI Press (1997) 291–316

[4] Florescu, D., Levy, A., and Mendelzon, A.: Database Techniques for the World-Wide Web: A Survey. SIGMOD Record, 27(3) (1998)

[5] Friedman, M., Levy, A. and Millstein. T.: Navigational Plans for Data Integration. AAAI-99 (1999) 67–73

[6] Genesereth, M.: An Agent-Based Framework for Interoperability. Bradshaw, J.M. (ed.): Software Agents. AAAI Press (1997) 315–345

[7] Hearst, M.: Information Integration. IEEE Intelligent Systems 13(5) (1998) 12–24

[8] Hsu, J.Y. and Yih, W.: Template-based Information Mining from HTML Documents. AAAI-97 (1997) 256–262

[9] Kitamura, Y., Noda, T., and Tatsumi, S.: Single-agent and Multi-agent Approaches to WWW Information Integration. Ishida, T. (Ed.): Multiagent Platforms, Lecture Notes in Artificial Intelligence, Vol. 1599, Springer-Verlag. (1999) 133–147

[10] Klusch, M. (Ed.): Intelligent Information Agents. Springer-Verlag (1999)

[11] Knoblock, C.A., Minton, S., Ambite, J.L., Ashish, N., Modi, P.J., Muslea, I., Philpot, A.G., and Tejada, S.: Modeling Web Sources for Information Integration. AAAI-98 (1998) 211–218

[12] Kuokka, D. and Harada, L.: Integrating Information via Matchmaking. Journal of Intelligent Information Systems 6 (1996) 261–279

[13] Kwok, C.T. and Weld, D.S.: Planning to Gather Information. AAAI-96 (1996) 32–39

[14] Russell, S.J. and Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice-Hall, Inc. (1995) 844

[15] Wiederhold, G.: Mediators in the Architecture of Future Information Systems. IEEE Computer, 25(3) (1992) 38–49

[16] Wiederhold, G. (Ed.): Intelligent Integration of Information. Kluwer Academic Publishers (1996)

[17] Wiederhold, G. and Genesereth, M.: The Conceptual Basis for Mediation Services. IEEE Expert, 12(5) (1997) 38–47

[18] Zhuge, Y., Garcia-Molina, H., Hammer, J., and Widom, J.: View Maintenance in a Warehousing Environment. SIGMOD-95 (1995).