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　知哉[††]　　辰巳　昭治[†]

An Intelligent Information Gathering Method 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情報基盤の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手法を提案する．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，適切な情報源アクセスと解の提示ができる．その評価としては，パラメータ変更可能な人工的な情報統合問題と，航空機の空席照会サービスという実世界の情報統合問題を対象に，従来の FIFO 型情報収集手法と比較して，提案手法の有効性を示した．

キーワード　情報エージェント，WWW，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，その利用価値をさらに高めることができる．

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を用いて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に対して，それを含む WWW ページの一覧を動的に作成して表示してくれる．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自動的に組み合わせて，利用者のクエリに答え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サービスが考えられる．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 を介して提供している．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 WWW サイトに個別に検索する必要がある．また乗り継ぎ情報が必要な場合にも同一の WWW サイトに対して繰り返し検索する必要が生じる．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ような空席照会サービスを実現することができれば，利用者をこのようなわずらわしさから解放することができる．

しかしながら以上のような WWW 情報統合を実現するには以下の問題点がある．

(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

[†] 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

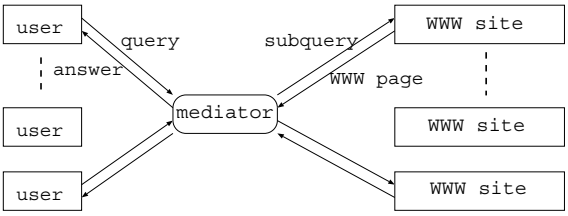

図 1　メディエータによる情報統合
Fig. 1　Infromation integration through a mediator.

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 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，過去に収集し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，必要であれば，利用者の代わりに複数の 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，一度得た情報はメディエータ内にキャッシュされる．しかしながら，情報源が頻繁に更新される場合には，一度キャッシュした情報であっても，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は重要な研究課題となる．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を提案する．

2.1 事　　実

メディエータは WWW ページを収拾し，そこから必要な情報を抽出する．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 2 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 3 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，それをキャッシュする機能をもつと仮定している．

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

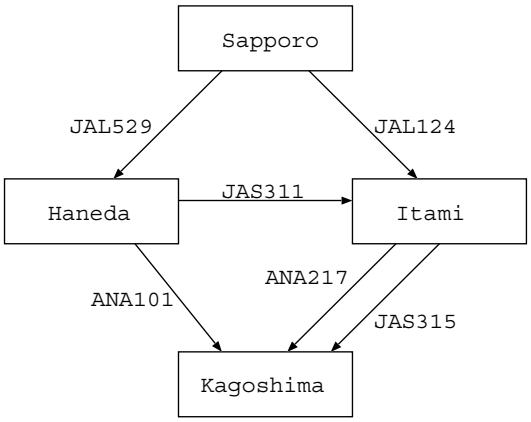

図 2　飛行経路
Fig. 2　Flight connection.

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 3　空席照会システムにおける事実
Fig. 3　Facts for flight information service.

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の信頼度は長時間一定のままである．

信頼度関数を本稿では

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

として近似している．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．

また，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見なすことができる．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見なせる．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ぼられる．

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 的にルール表現できる．

$(R1)$ 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．

例えば，空席照会システムにおける解導出は，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 3 で示されるキャッシュされた事実から，図 4 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 4 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

メディエータはキャッシュしている事実を用いて解を導出するが，解の導出に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

解はメディエータがキャッシュしている事実を用いて導出するが，キャッシュされてからの長時間経過している場合は，その解の信頼度は低下している．解の信頼度を向上させるには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に効率よく行う必要がある．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，これらのスコアを用いて，再検索を行う事実や利用者に提示する解を決定する．図 5 にそのアルゴリズムを示す．

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，できるだけ正しい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，再検索すべき事実を選択する (3 行目)．それは各解 A_i に対して，再検索スコア

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

を計算し，そのランク付けを行う．静的な質が大きく，信頼度の低い解は再検索スコアが大きくなるが，そのような解はそれを構成する事実を再検索すれば，信頼

(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 4　query(Sapporo,Kagoshima,1999/12/16,0600,1600) に対する解の集合
Fig. 4　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 the selected WWW page and update f_m and r(f_m) }
6: sort answers by P(·)
7: return the best answer to the user
```

図 5　情報収集アルゴリズム
Fig. 5　Information gathering algorithm.

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再検索されたとしても，質の高い解には至らないので再検索スコアは低くなる．再検索スコア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)，WWW サイトから関係するページを再検索し，事実の更新を行い，その信頼度を 1 にする (5 行目)．ここで信頼度を 1 にすることにより，同一の事実を何度も再検索することを防いでいる．

再検索の許容時間が過ぎると，メディエータは解を提示スコア $P(A_i)$ 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再検索されるかを示そう．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 4 参照) が，図 3 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 回目の再検索の際には 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) が再検索される．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．この時点で再検索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，与えられたクエリに対して，再検索の回数に応じて，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

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．

(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．

表 1　事実の再検索に応じた各関数値の変化
Table 1　Changes of function values according to the number of reloads

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

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．これは提案手法において，再検索スコアを

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

とした場合と等価である．

本稿ではまず，提案手法の一般的な性質を明らかにするために，パラメータの変更が可能な人工的な情報統合問題を対象に評価を行った．さらに，実世界での有効性を示すために，航空機の空席照会サービスメディエータを構築し，その評価を行った．

3.1　人工的情報統合問題での評価

メディエータが扱う問題領域によって，事実の更新頻度や，解を構成する静的事実や動的事実の割合は異なる．そこでこのような要因が提案手法の性能にどのような影響を及ぼすかを人工的な情報統合問題を用いて評価を行った．

問題の解は 10 個存在し，それぞれ 5 個の静的な事実と 5 個の動的な事実から構成されている．事実 f のとる値 $d(f)$ は 0 または 1 のいずれかである．動的な事実の更新間隔 u は正規分布 $N(u,1)$ に従い，すべての事実に共通とした．情報源へのアクセス時間はすべて一定とし，1 回の情報源アクセスで，1 つの事実のみの再検索が可能である．解 A の質 $Q(A)$ は，それを構成する事実の中で 1 をとるものの数とする．したがってここでは解の質は静的な質と動的な質の和として定義される．最後に，動的な事実 f_d の信頼度関数 $r(f_d)$ は，

$$r(f_d) = \frac{1}{1 + \frac{t}{u}}$$

とする．ここで，t はキャッシュしてからの経過時間，u は動的な事実の平均更新間隔である．

3.1.1　情報源の更新頻度の影響

問題に依存する条件としては情報源の更新頻度と利用者のアクセス頻度が挙げられる．ここでは利用者のアクセス間隔を 1 時間に固定し，事実の平均更新間隔 u を 1 時間，6 時間，1 週間に変化させて性能を評価した．

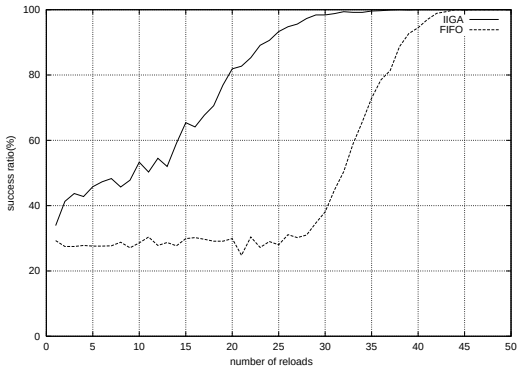

図 6　平均更新間隔 1 時間における性能
Fig. 6　Performance when mean update interval is 1 hour

まず，事実の平均更新間隔を 1 時間にしたときの結果を図 6 に示す．横軸が再検索回数，縦軸が最適解の提示確率 (成功率) である．解が 10 種類で，それぞれの 5 つの動的な事実を用いているので，動的な事実の再検索を 50 回行うと必ず最適解が得られることになる．利用者のアクセス間隔は 1 時間，事実の更新間隔も 1 時間なので，次の利用者がアクセスするまでに情報源の事実の大部分が更新されており，キャッシュはほとんど有効でない場合であるといえる．そのため，FIFO 法のようにキャッシュされた事実の古さだけを基準に再検索する事実を決定しても，ほとんど効果がない．一方，提案手法 (IIGA) では，静的な質を考慮することで，より重要と思われる事実を優先して検索することができ，よい性能を示している．

次に，事実の平均更新間隔を 6 時間にしたときの結果を図 7 に示す．更新時間が 1 時間の場合と比べて，どちらの方式も性能が良くなっている．これは，キャッシュされている事実がある程度信頼できるためである．特に FIFO 法の性能が改善されており，提案方式との性能差は更新間隔が 1 時間のときよりも縮まっているのが分かる．

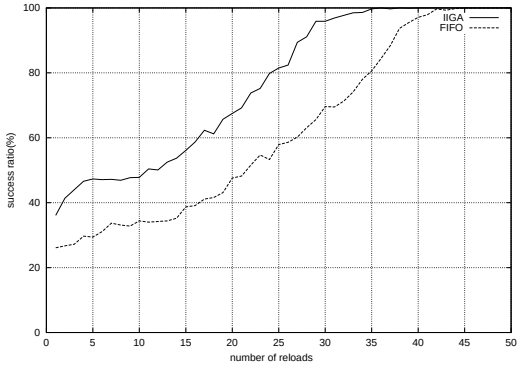

図 7　平均更新間隔 6 時間における性能
Fig. 7　Performance when mean update interval is 6 hours

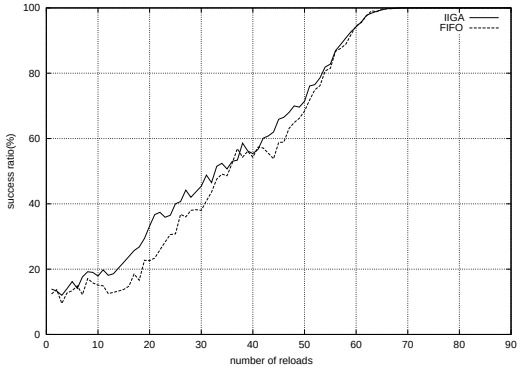

図 9　静的な事実 1 個，動的な事実 9 個の場合の性能
Fig. 9　Performance for solution consisting of 1 static fact and 9 dynamic facts

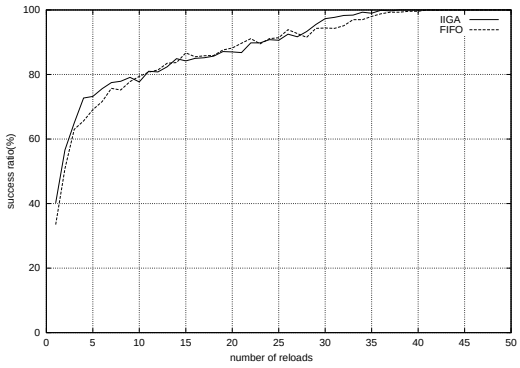

図 8　平均更新間隔 1 週間における性能
Fig. 8　Performance when mean update interval is 1 week

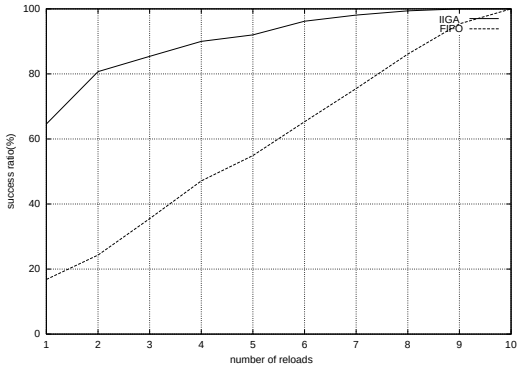

図 10　静的な事実 9 個，動的な事実 1 個の時の性能
Fig. 10　Performance for solution consisting of 9 static facts and 1 dynamic fact

最後に，事実の更新間隔を 1 週間にしたときの結果を図 8 に示す．先の場合よりも事実の更新間隔が長くなったため，キャッシュされている事実の多くが信頼できる状態にある．そのため，再検索の必要性が少なく，どちらの方式も同等の性能を示している．

これらの実験により，事実の更新が頻繁に行われる程，提案手法が有効であることがわかる．

3.1.2　静的な事実と動的な事実の割合の影響

解は静的な事実と動的な事実から構成されるが，その割合によって性能がどのように変わるかを調べた．ここでは解が静的な事実 1 個と動的な事実 9 個から構成される場合と，静的な事実 9 個と動的な事実 1 個から構成される場合を用い，それぞれの結果を図 9 と図 10 に示す．なお，動的な事実の平均更新間隔は 6 時間で固定している．

静的な事実の割合が少ない場合は，提案手法と FIFO 法の性能差はわずかである．これは，解の導出に用いる静的な事実が減少したので，静的な質による解の適切な評価が十分行われなくなったためである．

逆に，図 10 に示すように，静的な事実の占める割合が大きくなると，静的な質を用いて再検索する事実を決定する提案手法は，検索回数が少なくても良い性能を示すようになる

3.2　空席照会サービスでの評価

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

(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

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 11 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 法は約 18 回のアクセスが必要であることがわかる．これは FIFO 法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．

3.3 考　　察

ここまで人工的な情報統合問題と航空機の空席照会サービスという実世界の情報統合問題を通して提案した知的情報収集手法の性質を示してきた．評価実験を通して提案手法の有効性に関して明らかになったことは以下の 2 点である．

- 利用者のアクセス頻度と情報源の更新頻度の関係は性能に影響を与える．特に，情報源の更新頻度が大きい場合はキャッシュの信頼度が低下するので，適切に再検索を行う必要があり，FIFO 法に比べて提案手法の有効性は高い．
- 解を構成する事実において，静的な事実と動的な事実の割合は性能に影響を与える．例えば，静的な事実だけからなるような解は事実が更新する必要がないので，提案手法を用いる意味がない．また，動的な事実のみからなる場合は，信頼度のみを用いて事実の検索を行うことになるので，FIFO 法と同等の振舞いをすることになる．提案手法における再検索は静的な質を考慮して行われるので，評価実験からも示されるように，静的な事実の割合の多い方がより有効である．

さて，ここでは評価実験では議論されなった要因として，信頼度と解の質が提案手法にどのような影響を与えるかについて考察を加えておく．

事実の信頼度の推測も性能を左右させる重要な要因である．例えば，実際の信頼度よりも，推測した信頼度が早く減少するような場合は，再検索する必要がないのにもかかわらず，無駄な再検索を行ってしまう可能性が大きくなる．逆に，推測した信頼度が実際のものよりも遅く減少する場合には，事実がすでに更新されているにもかかわらず再検索が行われず，誤った解を導き出す可能性が高くなる．したがって適切に事実の信頼度を推測することが重要になる．

最後に，解の質の推測も性能に影響を与えることが予想される．すなわちその推測値が利用者の求めるものをより正確に表しているほど，利用者に対して適切な解を提示することができる．提案手法では，動的な事実は再検索するまでその値が確定しないことから，静的な質のみを用いて再検索すべき事実の選択を行っ

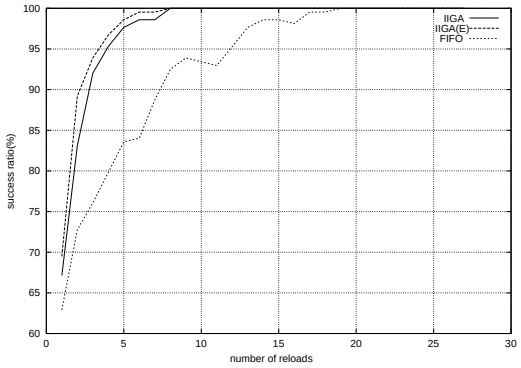

図 11　空席照会サービスにおける性能
Fig. 11　Performance in flight information service

(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．

ていたが，その期待値を求めてそれを再検索スコアに反映させることも可能であると考えられる．そこで空席照会システムにおいて，動的な事実である空席状況 f の期待値を以下のように定義した．

$$e(f) = \begin{cases} c(f) & f \text{が空席のとき} \\ 1 - c(f) & f \text{が満席のとき} \end{cases}$$

ここで，

$$c(f) = 0.5 + \frac{r(f)}{2}$$

である．したがって，信頼度が 1 から 0 に減少するにつれて，$e(f)$ の値は f が空席の場合は 1 から 0.5 へ，f が満席の場合は 0 から 0.5 へ変化することになる．

そこで解 A の質を，その解を構成する動的な事実を $f_1, \cdots, f_n$ としたとき，

$$Q_E(A) = Q_S(A) \times e(f_1) \times \cdots \times e(f_n)$$

として評価実験を行った．その結果は図 11 において IIGA(E) として示されており，わずかではあるが，期待値を用いる方が優れた性能を示している．これは，再検索する事実の決定は，動的な事実も考慮した方がより正確に行えることを示している．

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．

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．本稿ではキャッシュ内の事実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 法では事実の信頼度のみしか考慮していなかったのに対して，提案手法では解の質も考慮している．そのことにより FIFO 法よりも優れた性能を示すことを人工的な情報統合問題と，空席照会サービスという実世界問題に応用することで示した．本稿では具体的な応用事例として航空機の空席照会サービスを取り上げたが，提案手法は一般的に，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，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] S. Adali, K.S. Candan, Y. Papakonstantinou, and V.S. Subrahmanian, "Query caching and optimization in distributed miator systems," Proc. SIGMOD-96, pp.137–148, 1996.

[2] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom, "The TSIMMIS project: integration of heterogeneous information sources," Proc. IPSJ Conference, pp.7–18, 1994.

[3] T. Finin, Y. Labrou, and J. Mayfield, "KQML as an agent communication language," in Software Agents, ed. J.M. Bradshow, pp.291–316, AAAI Press, 1997.

[4] D. Florescu, A. Levy, and A. Mendelzon, "Database techniques for the world-wide web: a survey," SIGMOD Record, vol.27, no.3, pp.59–74, 1998.

[5] M. Friedman, A. Levy, and T. Millstein, "Navigational plans for data integration," Proc. AAAI-99, pp.67–73, 1999.

[6] M. Genesereth, "An agent-based framework for interoperability," Software Agents, J.M. Bradshaw, ed., AAAI Press, pp.315–345, 1997.

[7] M. Hearst, "Information integration," IEEE Intelligent Systems, vol.13, no.5, pp.12–24, 1998.

[8] J.Y. Hsu and W. Yih, "Template-based information mining from HTML documents," Proc. AAAI-97, pp.256–262, 1997.

[9] Y. Kitamura, T. Noda, and S. Tatsumi, "Single-agent and multi-agent approaches to WWW information integration," Multiagent Platforms, ed. T. Ishida, Lecture Notes in Artificial Intelligence, Vol. 1599, Springer-Verlag, pp.133–147, 1999.

[10] M. Klusch, ed., Intelligent Information Agents, Springer-Verlag, 1999.

[11] C.A. Knoblock, S. Minton, J.L. Ambite, N. Ashish, P.J. Modi, I. Muslea, A.G. Philpot, and S. Tejada, "Modeling web sources for information integration," Proc. AAAI-98, pp.211–218, 1998.

[12] D. Kuokka and L. Harada, "Integrating information via matchmaking," Journal of Intelligent Information Systems, vol.6 pp.261–279, 1996.

[13] C.T. Kwok and D.S. Weld, "Planning to Gather Information," Proc. AAAI-96, pp.32–39, 1996.

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

[15] G. Wiederhold, "Mediators in the architecture of future information systems," IEEE Computer, vol.25, no.3, pp.38–49, 1992.

[16] G. Wiederhold, ed., Intelligent Integration of Information, Kluwer Academic Publishers, 1996.

[17] G. Wiederhold and M. Genesereth, "The conceptual basis for mediation services," IEEE Expert, vol.12, no.5, pp.38–47, 1997.

[18] Y. Zhuge, H. Garcia-Molina, J. Hammer, and J. Widom, "View maintenance in a warehousing environment," Proc. SIGMOD-95, pp.316–327, 1995.

（平成 x 年 xx 月 xx 日受付）

北村　泰彦　（正員）

1983 大阪大学基礎工学部情報工学科卒. 1988 同大学院博士課程了. 工学博士. 1988 大阪市立大学工学部電気工学科助手. 現在, 同情報工学科助教授. 分散人工知能, ヒューリスティック探索, WWW 情報統合の研究に従事. IEEE, AAAI, ACM, 人工知能学会, 情報処理学会, ソフトウェア科学会等の会員

野田　知哉

1998 大阪市立大学工学部情報工学科卒. 2000 同大学院修士課程了. 2000NTT コムウェア (株) 入社. 在学中, WWW 情報統合の研究に従事.

辰巳　昭治　（正員）

1970 大阪大学工学部通信工学科卒. 1972 同大学院修士課程了. 1972 川崎重工業入社. 1978 大阪大学大学院博士課程了. 工学博士. 豊橋技科大学を経て, 現在大阪市立大学工学部情報工学科教授. 統計的パターン認識, 意思決定問題, 画像処理用並列プロセッサの開発, VLSI 向き相互結合網の構成法などの研究に従事. 電子情報通信学会, 情報処理学会, ソフトウェア科学会, IEEE 等の会員.