アクティブ情報収集システムに関する検討

A Thought on Active Information Gathering Systems

北村泰彦
kitamura@info.eng.osaka-cu.ac.jp
大阪市立大学大学院工学研究科

概要

An active information gathering system efficiently collects information from a number of frequently updating information sources on the Internet, considering the quality and cost of information gathering, to meet demands from human users or active mining systems. In this paper, we summarize functions required for active information gathering systems and related works. Then we propose a system called Intelligent Ticker. Intelligent Ticker consists of multiple information gathering modules and an information integration module. An information gathering module produces Tickers based on the difference between an updated Web page and the original one. The information integration module integrates multiple Tickers by using a Ticker Integration Template to assist the user in his decision making or problem solving.

Keywords: WWW, Active Information Gathering, Information Integration, Ticker

1. はじめに

インターネット(Internet)は我々の生活を支えるインフラストラクチャの一つとして急速に社会に浸透しつつある．インターネットをベースとしたサービスの中でも WWW (World Wide Web)はその中でも最も人気が高く，学術研究，電子商取引，個人やグループなどによる情報発信など，さまざまな目的のために利用されている．WWW はいまや情報量の見地からは地球規模の知識ベースを構築しているといっても過言ではないであろう．

WWW システムの特徴は，従来の分散データベースシステムと異なり，情報源がボトムアップに構築されるところにある．情報発信者はコンピュータをインターネットに接続し，Web サーバを立ち上げるだけで，即座に世界に向けた情報発信が可能になる．このように WWW システムは，集中的な管理機構無しに，膨大な数の情報源が自律分散的に連携しているシステムであるといえる．

一方で，一利用者の観点からは有用な情報がネットワークの中に埋没してしまい，容易に見つけ出すことができないという問題も引き起こしている．これに対処するための一時的な解決法として検索エンジンが開発されている．しかしながら検索エンジンはキーワード入力に対して，それを含む Web ページの集合を出力するだけである．中には，膨大な数の Web ページを出力したり，また多数の関係のない Web ページを含むようなものも存在する．今後は得られた Web ページから有用な情報を自動的に発見するようなデータマイニングシステムが望まれる．

しかしながら，このようなデータマイニングシステムの実現するためには以下のような Web 情報源の特徴を考慮する必要がある．

- Web 情報の記述形式は非定型である．Web 情報源は一般にそれぞれが URL(Universal Resource Locator)によって指定可能な Web ページの集合と見なすことができる．Web ページは一般に HTML (Hyper Text Mark-up Language)により記述されていることが多いが，HTML では Web ページをブラウザで表示する際に必要な視覚的な構造を表現することが可能であっても，Web により記述されている情報の意味的な構造を記述することは困難である．この問題への対処とし

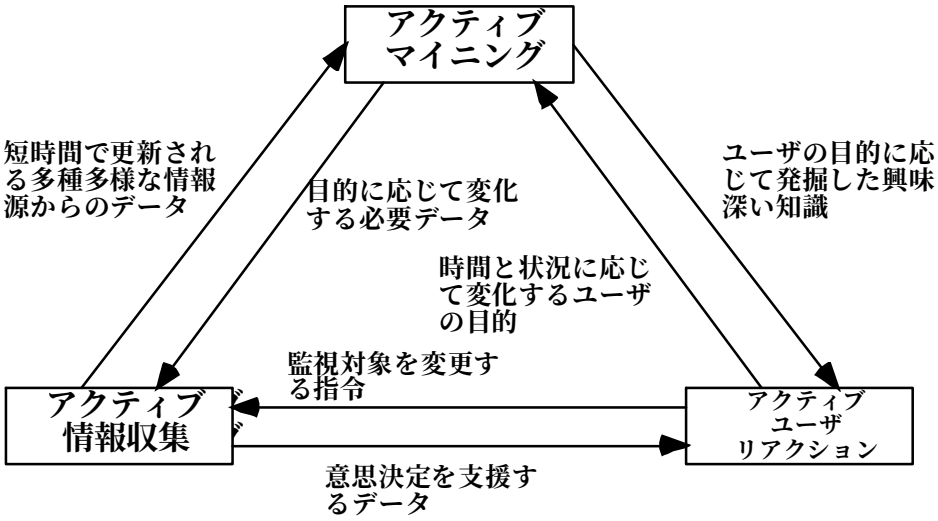

図 1：アクティブマイニングシステムの構成図

ては意味的構造も記述可能とする次世代の Web ページ記述言語である XML[1]や Web 情報の機械的処理を目的とした Semantic Web[2]の研究や導入が行われている.

- Web 情報源に蓄積されている情報は日々, 急速な勢いで増加している. 多くの情報源が 1 日に数回の情報更新を行っている. また株価情報や道路情報などは数分毎に情報更新を行っているものも少なくない. データマイニングシステムが情報収集し, 何らかの情報を発見したとしても, その情報がすでに古いものであれば, それは利用者にとって有用であるとはいえない.

本論文では以上のような Web 情報源の特徴を考慮しながら, 膨大な数の動的な情報源の中から利用者にとって適切な情報源を選択し, その中から有用な情報を発見するのがアクティブマイニングシステムについて議論する. アクティブマイニングシステムの構成は図 1 に示される. [3]

- アクティブ情報収集モジュールは動的で大規模なインターネット情報源からアクティブマイニングモジュールや利用者に対して必要な情報を監視・提供する役割を果たす.
- アクティブマイニングモジュールはアクティブ情報収集モジュールにより収集された情報を解析し, 利用者にとって有用な情報を発見する.
- アクティブユーザリアクションモジュールは利用者とのインタフェースの役割を果たし, 利用者の要求に変化が生じた場合はそれをアクティブ情報収集モジュールやアクティブマイニングモジュールに伝達する.

以上のような三つのモジュールが互いに連携しあうことによりアクティブマイニングが達成される.
本論文ではこの中でアクティブ情報収集に焦点を絞り, 2 節ではアクティブ情報収集システムに要求される機能と関連研究について述べる. 3 節では Web ページの差分情報を基づくアクティブ情報収集システム Intelligent Ticker の構想について述べ, 4 節でまとめとする.

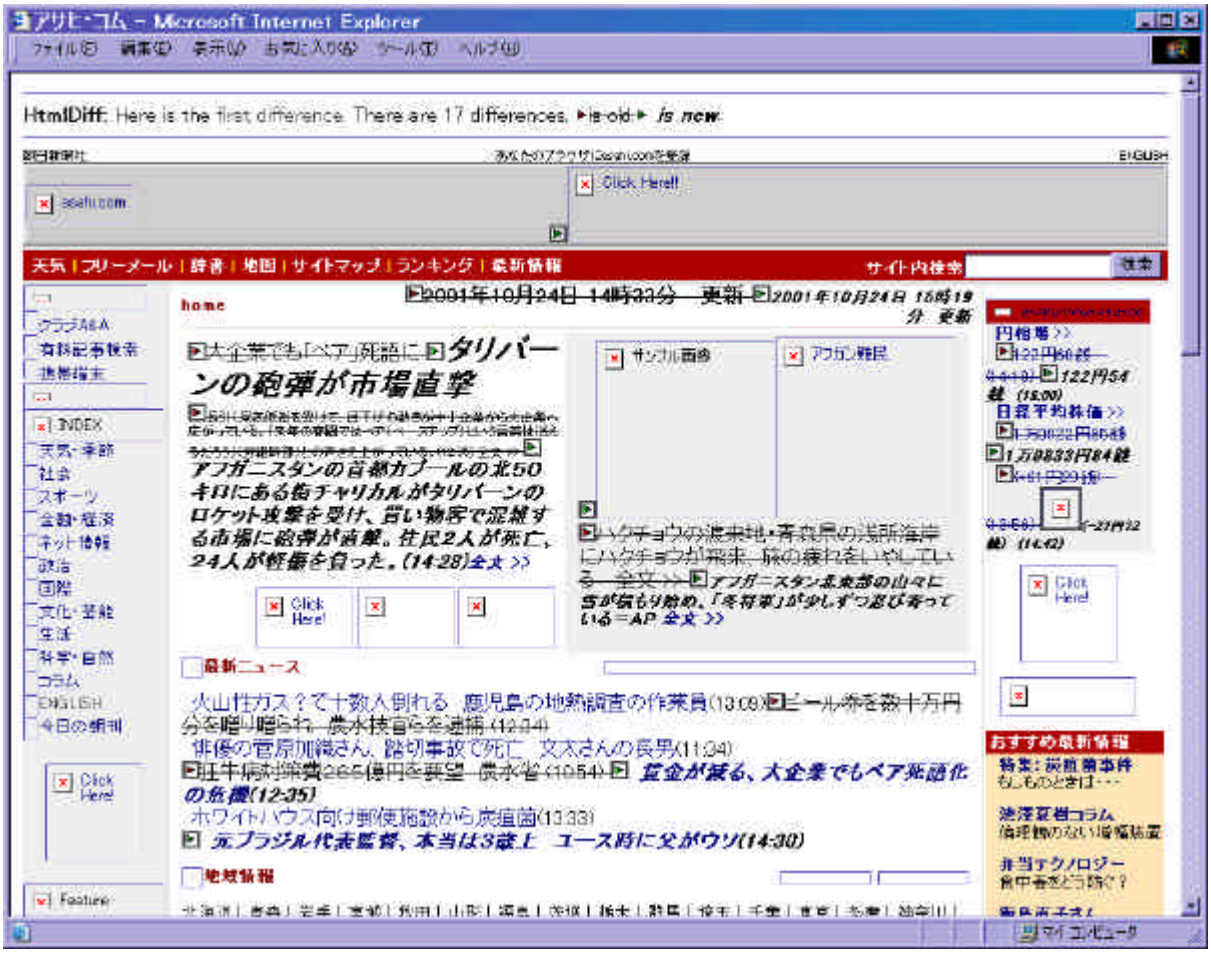

図 2：HtmlDiff の出力画面

2. アクティブ情報収集システム

アクティブ情報収集システムはインターネット上に存在する動的に変化する Web 情報源から利用者の要求を満たす情報を効率よく収集し，さらにその変化を監視するシステムである．アクティブ情報収集システムに求められる機能は以下のようにまとめられる．

情報監視機構

Web 情報源の特徴の一つは情報源が頻繁に更新されることである．更新の頻度は情報源により異なるが，大学のホームページのように 1 週間に 1 度程度しか更新が行われないようなものから，新聞社のホームページのように 1 時間に数回の更新が行われるようなものもある．さらには，株価，スポーツ中継，オークション，道路情報などの情報提供サイトでは数分に一度の割合で更新が行われるものもある．

利用者はこのように頻繁に更新される情報源から効率よく情報収集を行うことを望んでおり，そのような要求に応えるいくつかのシステムが開発されている．

AT&T で開発された AIDE (AT&T Internet Difference Enigne)[4]は WWW の変化を監視し，その違いを表示するシステムである．この中で二つの Web ページを比較し，その違いを表示する HtmlDiff[a]と呼ばれるモジュールが開発され公開されている．またその改良を行い，Java により実装したものとして TopBlend[5]がある．

Lawrence Livermore National Laboratory で開発されている DataFoundry[b]は情報源の変化を発見し，データウェアハウスのメンテナンスを行うシステムである．[6,7]ここでは科学データ源におけるデータベーススキーマをグラフ表現し，データとスキーマの変化の検

---

[a] http://www.research.att.com/~douglis/aide/
[b] http://www.llnl.gov/casc/datafoundry/index.html

知する．従来，科学データベースにおいてスキーマ変更は頻繁に行われるが，それを手作業で行うにはコストがかかっていた．情報源を定期的に監視し，自動的にスキーマを変更しようとする試みである．
また INRIA では XML ベースのデータウェアハウスのためのデータ監視システム Xyleme[c]が開発されている．[8]
これ以外にも情報源監視を連続的なクエリ（continuous query）とみなすデータベースシステムからのアプローチとして Oregon Graduate Institute の CONQUER[9]や University of Wisconsin の NiagaraCQ[10]と呼ばれるシステムも開発されている．

差分表示機構
Web 情報源の変化が検知されたとき，利用者はその変化が知らされるだけでなく，どのように変化したかを分かりやすく知りたいという要求がある．このような目的にために前節で述べた HtmlDiff システムでは図２のように，二つの Web ページの違いを際立たせるために過去のデータには取り消し線を，新しいデータはイタリックで表示させるようにしている．

評価機構
Web 上には同様の情報を扱う情報源が多数存在する．例えば，新聞社のホームページなどは多数あり，同様の情報を発信しているといえるが，その視点や頻度はそれぞれ異なっている．利用者はより早く，より有用な情報を入手したいと望むであろう．これは複数の情報源を監視し，更新の早さや量を比較することにより，情報源の近似的な評価を行うことが可能になると思われる．

統合機構
複数の Web 情報源から得られる情報を統合することはそれぞれの情報源の付加価値を高めることになる．[11]例えば，新聞サイトから東海道新幹線が運休するニュースを入手した東京にいる旅行者が，航空会社のサイトから羽田空港発の航空便の空席情報を直ちに入手できるならば有意義である．また，同種の情報源からの情報を組み合わせることも有用である．例えば，同じ話題のニュースであっても，それが多くの情報サイトで取り上げられているとすれば，そのニュースがより重要であることが分かる．
このような情報統合機構を実現する場合には，情報収集の質とコストを考慮する必要がある．[12]例えば，情報の質を収集した情報源の数で評価するとするならば，より高い質の情報を得るためには，より多くの情報源から情報収集する必要があり，より多くのコストが必要になる．したがって一般には情報収集の質とコストはトレードオフの関係にあるといってよいであろう．質とコストをうまくバランスをとりながら情報収集するためにはそのためのプランニング機構が必要になる．このような目的で Massachusetts 大学では BIG と呼ばれる情報収集エージェントが開発されている．[13]

3. アクティブ情報収集・統合システム Intelligent Ticker の構想

頻繁に更新される情報源からの情報収集を考えた場合，一般に更新されるそれぞれの情報の量はそれほど大きくはない．例えば，新聞社のサイトにおいてそのトップページは数分の単位で頻繁に更新されるが，更新される量は Web ページ中の数行である．テレビ等で行われるニュース速報にしても数行のテキストが画面の上部に表示されるだけである．
我々はこのように少量で速報性のある情報オブジェクトを Ticker と呼び，それらを Web 情報源から収集し，統合することで利用者の意思決定や問題解決を支援する Intelligent Ticker システムを提案する．
Intelligent Ticker は図3に示すように情報収集部と情報統合部から構成される．情報収集部は指定された Web 情報源に対して，その変化を監視し，その変化が存在した場合は，その差分のみを Ticker と呼ばれる情報オブジェクトを生成する．

情報統合部は複数の情報収集部で生成された Ticker を選択，組み合わせることにより利用者の意思決定や問題解決を支援する．利用者は情報統合部で得られた Ticker を直接表示させることも可能である．

[c] http://www.xyleme.com/index.jsp

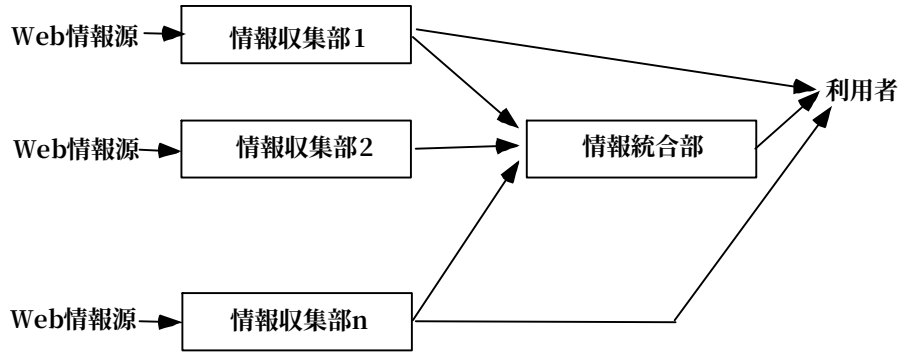

図３：Intelligent Ticker の構成図

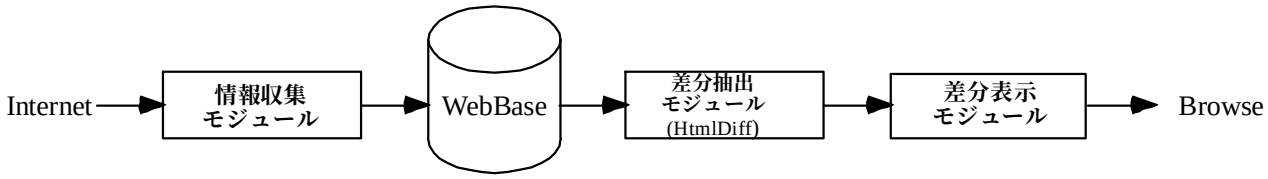

図 4：情報収集部の構成

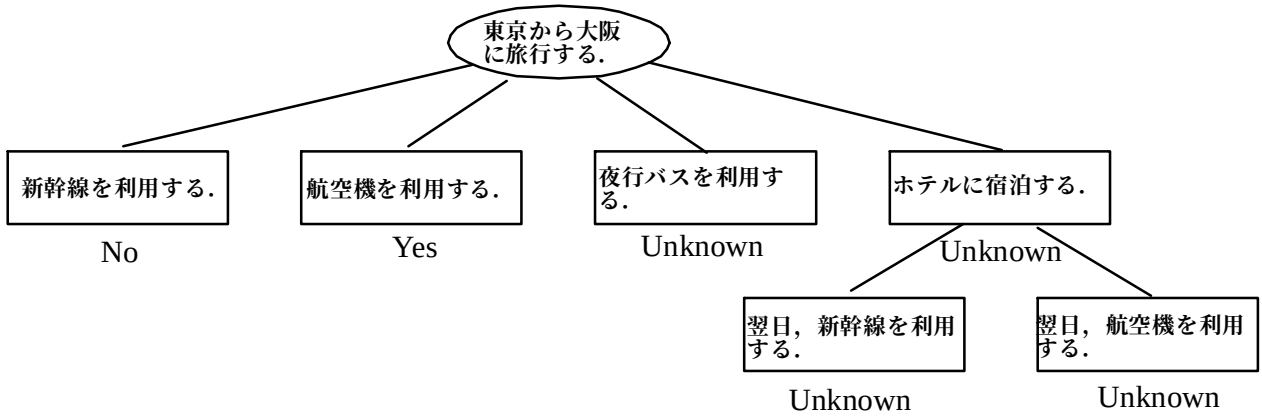

図 ５：Ticker Integration Template

情報収集部

情報収集部の構成は図４のようになっている．情報収集モジュールはインターネット上の情報源から定期的に Web ページをフェッチし，WebBase に保存する．差分抽出モジュールは利用者から WebBase に格納されている二つの Web ページの指定に対してその違いを抽出する．このモジュールは先述の HtmlDiff を利用することができる．HtmlDiff はもともとの Web ページにタグを埋め込むことによりその違いを表示しているが，差分表示モジュールでは図 5 のように，変化した部分だけを利用者に対して表示する．

差分表示モジュールでは差分抽出モジュールにより抽出された差分を表示するが,差分のみを表示するだけでは，その文脈が取り去られてしまい差分の意味が理解しにくくなってしまうことが考えられる．そこで Web ページを構成する HTML 文書のタグの入れ子構造を解析し，差分の上位概念を示す構造は残すようにしている．すなわち図５の例では，ニュースの差分だけでなく，そのジャンルを表す「最新ニュース」,「社会」,「経済」といったヘッダ情報も表示されている．

情報収集部は情報源の変化に対して Ticker を生成する．Ticker は以下の要素から構成される．

- **オブジェクト：更新された情報の断片そのもの．テキストやハイパーリンクなどにより表現され**

る.

- タイムスタンプ：更新された時刻.
- ロケーション：更新された情報の URL.
- コンテキスト：更新された情報の文脈.

情報収集モジュールでは情報源の更新頻度を考慮した情報アクセスが望まれる．すなわち，例えば，1 日に 1 回しか更新されないことが分かっている情報源に対して, 1 時間ごとに情報収集したとしても無駄である．したがって情報収集の操作を行いながら，情報源の更新頻度を学習し，それに応じて情報収集の間隔を変化させるような適応的な機能が必要になるであろう.

情報統合部

情報収集部は一つの情報源を監視し，変化が生じたときに Ticker を発生するシステムである. 情報統合部は複数の情報収集部で発生する Ticker を選択し, 統合する.

Ticker の統合には図に示されるような TIT (Ticker Integration Template)が用いられる．この図では東京から大阪に旅行する場合のプラン候補が示されている．現在情報統合部は新幹線情報に関する Ticker と航空便情報に関する Ticker を選択しており, 現在新幹線は満席で利用不可能，航空機は空席があり利用可能であるとする．ここで航空機が結構になったことを通知する Ticker を受け取ると，新たに夜行バスに関する Ticker を収集し，それが利用可能でない場合は，ホテルと翌日の交通手段に関する情報を収集する．もちろん，情報収集の途中で新幹線に空席が生じれば，夜行バスやホテルに関する情報収集は停止してもよい．このように情報統合部では得られる Ticker の内容に応じて動的に情報収集の方法を変更してゆく.

4. まとめ

アクティブ情報収集システムに要求される機能についてまとめ, Intelligent Ticker と呼ばれるアクティブ情報収集システムの提案を行った.

謝辞

本研究は科学研究費補助金（特定領域研究(B)）「分散動的情報源からのアクティブ情報収集」（課題番号 13131209）によるものである.

参考文献

[1] Klein, M.: XML, RDF, and Relatives. IEEE Intelligent Systems 16:2, 26-28 (2001)

[2] Fensel, D., Musen, M.A.: The Semantic Web: A Brain for Humankind. IEEE Intelligent Systems 16:2, 24-25 (2001)

[3] 元田浩：情報洪水時代におけるアクティブマイニングの実現，科学研究費補助金「特定領域研究(B)」申請書 (2001)

[4] Douglis, F., Ball, T., Chen, Y.-F., Koutsofios, E.: The AT&T Internet Difference Engine: Tracking and Viewing Changes on the Web. World Wide Web, 1: 27-44 (1998)

[5] Chen, Y.-F., Douglis, F., Huang, H., Vo, K.-P.: TopBlend: An Efficient Implementation of HtmlDiff in Java. In WebNet'00 (2000)

[6] Adam, N., Adiwijaya, I., Critchlow, T., Musick, R.: Detecting Data and Schema Changes in Scientific Documents. In IEEE Advances in Digital Library (2000)

[7] Critchlow, T., Fidelis, K., Ganesh, M., Musick, R., Slezak, T.: DataFoundry: Information Management for Scientific Data. IEEE Trans Inf Technol Biomed, 4(1): 52-57 (2000)

[8] Nguyen, B., Abiteboul, S., Cobena, G., Preda, M.: Monitoring XML Data on the Web. In ACM SIGMOD (2001)

[9] Ling Liu, Calton Pu, Wei Tang, and Wei Han. Conquer: A continual query system for update monitoring in the www. International Journal of Computer Systems, Science and Engineering, 14(2): 99-112 (2000)

[10] Jianjun Chen, David DeWitt, Fend Tian, and Yuan Wang. NiagaraCQ: A scalable continous query system for the internet databases. ACM SIGMOD, 379 (2000).

[11] 山田誠二，村田剛志，北村泰彦．知的 Web 情報システム，人工知能学会誌，16(4):495-502 (2001)

[12] 北村泰彦，野田知哉，辰巳昭治．動的情報メディエータのための知的情報収集手法，電子情報通信学会論文誌 D-I, J84-D-I(8):1256-1265 (2001)

[13] Lesser, V., Horling, B., Klassner, F., Raja, A., Wagner, T., Zhang, S.X.: BIG: An agent for resource-bounded information gathering and decision making. Artificial Intelligence, 118(1-2): 197-244 (2000)