Institute of Electronics, Information, and Communication Engineers

D-8-22

WWW における複数インスタンス情報源のための wrapper 生成機構について

Wrapper Generating Mechanism
for Multiple Instance Sources on WWW

山田亮太　伊藤孝行　福田直樹　新谷虎松
Ryota Yamada　Takayuki Ito　Naoki Fukuta　Toramatsu Shintani
名古屋工業大学 知能情報システム学科
Dept. of Intelligence and Computer Science, Nagoya Institute of Technology

1. はじめに

近年，WWW を通して様々な情報を入手することが可能になった．WWW 上に存在する情報は膨大であるため，必要な情報の発見や選別が困難になるといった現象が見られる．この問題に対して，エージェントと呼ばれる自律的ソフトウェアを用いて WWW 利用者の支援を行う研究がある．WWW では多くの情報は HTML により記述された文書 (HTML 文書) として提供される．HTML は情報の表示形式の記述を目的としたマーク付け言語であるため，HTML 文書中で情報が意味内容に基づいて構造化されていることは少ない．エージェントにより情報を効果的に利用するためには，情報を意味内容に基づいて構造化する必要がある．情報源から意味内容に基づいて構造化された情報を抽出する仕組みとして，wrapper と呼ばれるものが存在する．wrapper は情報の提示形式を解析して人手で作成することができる．人手による wrapper の作成の負担を軽減するために，wrapper の自動的な生成を目指した研究が行われてきた．本研究では，複数インスタンス情報源から提供される HTML 文書に対する wrapper を，人手による HTML 記述の解析を行うことなく，半自動的に生成するための wrapper 生成機構を提案する．複数インスタンス情報源とは，一定の提示形式を用いて複数の異なる情報を提示する情報源である．WWW 上には CGI 等を用いて HTML 文書を自動生成してサービスを提供する Web サイトが多数存在する．このような Web サイトは複数インスタンス情報源である．自動生成される HTML 文書には冗長な情報が多い．本研究では冗長な情報を除去する wrapper の生成機構を提案することで WWW 利用者が真に必要とする情報の発見や選別を支援することを目指す．

2. 複数インスタンス情報源のための wrapper 生成

wrapper の構築には，情報源における情報の記述の構造の分析に基づいた，情報の記述に関する規則の発見が必要となる．HTML 文書を対象として考えた場合，(a)HTML 文書の数は非常に多い，(b) 新しい HTML 文書が頻繁に登場する，(c) 既存の HTML 文書の記述は頻繁に変化するという 3 つの理由より，手作業での wrapper の構築は現実的でなく，wrapper を可能な限り自動的に生成する機構が必要となる．

同じ複数インスタンス情報源から取得可能な HTML 文書には，情報の提示形式を規定する共通した記述が存在する．本稿では，この共通した記述を雛形と呼ぶ．雛形は情報量の低い情報を除去して情報量の高い情報のみを抽出するマスクとして，同じ複数インスタンス情報源の HTML 文書に適用することが可能である．本稿では，雛形を利用して情報を抽出し，情報源における情報の提示形式に関する知識を用いて，抽出した情報を構造化する手法を提案する．複数インスタンス情報源における雛形は，システムが自動的に獲得することが可能である．情報の構造化に用いる知識としては，情報源において情報が提示される順番などを用いる．この知識を構築する上で，利用者が HTML 記述を解析する必要は無い．提案手法を用いると，利用者は HTML 文書の解析を行うことなく，情報を構造的に抽出することが可能となる．

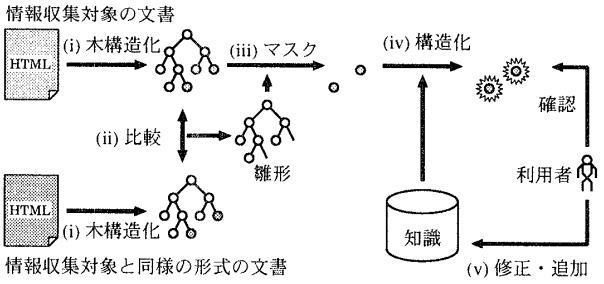

図 1: wrapper 生成の流れ

本機構における処理の流れを図 1 に示す．

以下，図 1 に示した (i) から (v) の処理について概説する．(i) 情報収集の対象とする HTML 文書および情報収集の対象とする HTML 文書と同様の形式で情報を提示する HTML 文書を木構造で表現する．(ii) 木構造に変換した文書を比較して雛形を取り出す．(iii) 木構造化された情報収集対象の HTML 文書に雛形をマスクとして適用し，情報を抽出する．(iv) 抽出した情報を構造的に利用することを可能にするために，予めシステムに知識として与えられた情報を利用して抽出した情報の構造化を行う．(v) 情報源における情報の提示形式に変更が生じ，予め与えられた知識では正しい構造化を行うことができなくなった場合，必要に応じて利用者がシステムに与えられた知識の修正あるいは補強を行う．

以上が本稿で提案する wrapper 生成の流れである．(i) から (v) の 5 つのステップを経て得られた雛形と知識を併用すると，情報収集の対象と同様の形式で情報を提示する HTML 文書から情報を構造的に収集することが可能になる．したがって，本 wrapper 生成機構により得られる雛形と知識を併せたものを wrapper であるとみなすことができる．

本 wrapper 生成機構を複数オンラインオークション入札支援システム *BiddingBot*[Ito *et al.*, 2000] に組み込み，情報収集を試みたところ，情報源で用いられる雛形が単一である際に極めて良好に機能することを確認することができた．

3. おわりに

本稿では，利用者による HTML 記述の解析を必要としない wrapper 生成機構を提案した．本機構では，雛形による情報抽出と抽出される情報に関する知識を用いることにより，複数インスタンス情報源のための wrapper を実現した．

今後の課題としては，現実的な複数インスタンス情報源において情報の提示形式に含まれる微妙な差異を吸収するような，質の高い wrapper 生成の実現が挙げられる．このためには，情報に関する知識の質の向上などが必要となる．

参考文献

[Ito *et al.*, 2000] Takayuki Ito, Naoki Fukuta, Toramatsu Shintani, and Katia Sycara, "*BiddingBot*: A Multiagent Support System for Cooperative Bidding in Multiple Auctions," In the Proceedings of the Fourth International Conference on Multi Agent Systems (ICMAS2000), 2000 (to appear as poster abstract).

NII-Electronic Library Service