B-31　オントロジを用いた情報配信システムの研究開発

Development of Information Notification Service using Ontology

日高 由布子　槌谷 一†
Yufuko Hidaka　Hajime Tsuchitani

1．はじめに

Web 上には様々な情報が氾濫している．そういった中からユーザが本当に欲しい情報を効率よく入手するといった作業は年々困難になってきている．近年 SemanticWeb に代表されるような機械により処理(理解)可能な Web 及びサービスを提供する技術開発が盛んになっている．なかでもオントロジに関する研究開発は W3C[1], DARPA[2], EU IST[3]といった団体で標準化が進められている．

本論文ではネットワーク上に散在する情報の中から的確な情報をユーザに配信するサービスの利便性の向上に焦点を当て，情報の key となる語彙に対しこのオントロジ技術を用いることで，既存の情報配信システムよりも情報収集効率がよく，また非構造的な既存の Web 上の情報から効果的に情報を取り出し，さらに，ユーザへ提供する際により利便性がよい情報に加工して情報を配信する情報配信システムの研究開発を行った．

2．オントロジを用いた情報配信システム

2.1　オントロジとは

オントロジは，ある領域の重要な概念の合意され，共有された形式的な記述で，領域中のオブジェクトのクラスを識別し，クラス間の階層構造を構成する．個々のクラスはプロパティによって特徴付けられ，クラス間，プロパティ間の関係を持つ．オントロジは問題領域の知識を記述する潜在的な用語を提供する．

2.2　情報配信システムとは

情報配信システムはユーザの代わりに自動的に情報を集め，それを配信するシステムである．現在稼動中の情報配信システムでは，ユーザがあらかじめあるカテゴリ，例えばニュース配信の場合には政治やスポーツといったカテゴリを登録しておくと，その登録にマッチしたニュースがユーザに配信されたり，ユーザのポータルページに表示されるといったシステムである．このようなカテゴリにより情報が分けられているシステムを Topic-Based Publish/Subscribe システムと呼ぶ．

既存のニュース配信などのシステムにおいてはこの Topic-Based Publish/Subscribe システムであるため，ユーザが得る情報は大きく分けられたカテゴリ内の情報全てになってしまい，ユーザの本当に知りたい情報の選択は情報配信後のユーザに委ねられている．

上述した Topic-Based Publish/Subscribe システムと異なるタイプのシステムで Content-Based の情報配信システムとして弊社の製品 WES (Websphere Everyplace Server) V2.1 のコンポーネントである INS (Intelligent Notification Services) があげられる．ユーザは「IBM の株価が 100 を超えたら通知して」というようなカテゴリ分けされていない興味を登録することが可能である．この場合情報源では IBM の現在の株価は 120 といった情報から key=IBM, stock=120 といった key/value ペアを作成し，INS のマッチングエンジンである Gryphon Broker[4]へとこれを渡す．すると key=IBM, stock>100 に該当するかどうかの判断が下され，該当している場合にはユーザへメール等で通知される仕組みである．

このシステムではユーザのニーズは Topic-Based のシステムよりも細かく反映されている．しかし，Gryphon Broker でのマッチングにおいて，例えば上述した例における key になる「IBM」という値が「アイ・ビー・エム」となっている場合には別なものであると認識し、情報が存在するにもかかわらず見落としてしまうといった事が起こる場合がある．各情報提供者が統一した名称を使用し情報を提供した場合には有効な情報配信システムであるが，使用する言葉に揺れが存在した時点で例えマッチしていたとしてもその情報は省かれてしまう．

またユーザが欲しい情報そのものは提供されていないが，提供されているデータよりユーザが算出できるようなデータがある場合も考えられる．株価の例でいうと，前日の値と当日の値のみが提供されているが，前日からどれくらい差異があるか提供されていなかった場合，ユーザはこれらの値から自身で差異を換算できるが，もしユーザが差異自体を興味ある項目としてシステムへ登録した場合には，この情報源からは情報は得られないと判断されてしまう．

そこで本論文ではこの INS をベースのサーバとして用い，これにオントロジを利用し，ユーザの興味ある情報をより多く収集できるようなシステムを考案した．そしてさらにオントロジを用いることを活かし，既存の情報からさらにユーザに有益な情報を作成することを目標とした．

2.3　アーキテクチャ

弊社のポータルサーバである WPS (Websphere Portal Server)でユーザは自身の興味を登録する[5]．本論文で登録したユーザの興味は「家賃総額が 9 万円以内で部屋の面積が 40 平米より広い部屋を表示して」という内容とする．登録されたユーザの興味は INS へ送られ，データベースにトリガーとして登録される．このトリガーと一致した情報がきた場合，ユーザへ通知される．

一方 Web 上にある情報を収集する Monitor agent により既存のコンテンツプロバイダーや Web サイトより情報は集められ，集めた情報から key/value ペアを作成し，INS の Gryphon Broker へ渡しユーザの興味とのマッチしているか調べ，マッチしている key/value が存在すると，ユーザへ delivery monitor を経由して情報が提供される仕組みである．

情報を収集する際と key/value ペアを作成する際に key に対してオントロジを用いることで，上位下位概念，関係処理を施し，言葉の揺れなどの形態素にまつわる処理を行う．

異なった語彙であるが同じ概念を表すものは等しいという関係をつけたり，ある概念同士を用い新たな別のものへ

† (株) 日本アイ・ビー・エム　ソフトウェア開発研究所

変換したりすることで，同じ意味でも既存のマッチングエンジンでは省かれてしまっていたものを拾い集めたり，収集した語彙に対する語彙間の関連付けにより，有益な情報に変換してユーザに提供することが可能となる．これらの詳しい例は以下の実装例で見ていくことにする．

2.4 研究成果の実装例

本論文において，我々は情報配信の源として 2 つの情報源に対する適応方法を考案した．1 つ目はあらかじめ形式がそろっている NITF[6]というニュース配信に用いられる形式をした情報がコンテンツプロバイダーから得られることを仮定したシステムへの適応で，2 つ目は Web 上にある HTML 形式で書かれた情報への適応である．

まず最初に NITF 形式をした情報への適応について述べることにする．形式は揃っていても情報提供者により同じ情報を表すのに語彙が違うことが多々ある．今回の適用例の賃貸情報を得るという仮定の下では 2 種類の NITF 形式の情報源を用意した．情報の中身の例として，1 つは<td class="専有面積">43</td>，もう 1 つの情報源では<td class="面積">48.6</td>といったような部屋の面積の情報を考える．ここから INS のマッチングエンジンへ渡す key/value を作成するのであるが、この例のように key が違う言葉で表されているが意味は同じであるといった語彙はマッチングエンジンでは同じ物とみなされず，情報が拾いきれないことになる．

このように形式は同じであるが語彙が異なるものに対し，

Class 専有面積 sameAs 面積
Class 面積 sameAs 専有面積

といったオントロジを用いることで 2 つの語彙は同じ概念を表すという関連付けを行うことが可能となるため，よりユーザのニーズに合った情報まで拾うことが可能となる．また面積の例では，単位が異なる場合，例えば「何畳」，「平米」と異なっている場合でも変換ルールを規定することによりユーザが比較可能な形式に変換できる．

さらに「家賃総計」は「家賃」＋「管理費」といった概念を以下のように定義し，以下のような定義の場合「家賃総計」を算出するようにルールを決めることで，元の情報源には無かった「家賃総計」といったものも提供でき，ユーザに対しさらに有益な情報へ変換することができる．

Class 家賃総計 subClassOf
 unionOf 家賃
 管理費

様々な領域に対応するオントロジを用意し，これを呼び出すことで色々な情報に対応することが可能となる．

次に 2 つ目の HTML 形式をした情報源から INS のマッチングエンジンへ渡す key/value ペアを作成する方法について述べることにする．SemanticWeb の標準化が進められている段階の現状においては，Web 上の HTML で書かれた情報については機械が処理できるメタデータはついていない．このような HTML から INS のマッチングエンジンへ渡す key/value ペアを作成することは困難であった．

そこで本論文では，HTML で記述された情報のうち特に表形式でかかれている場合，「table」タグに着目して情報部分の表を切り出し，切り出された表のテーブル内の値にオントロジを用いて推論を行い有益な情報を取り出した．

例をあげてみてみることにする．賃貸情報が 2 つの Web サイト上に表形式を用いて提供されているとする．NITF 形式の場合と同様に「専有面積」と「面積」は等しいといったオントロジや「所要時間」には「バス使用」や「徒歩」が存在するというオントロジ，「家賃総計」の算出，「初期費用」の算出(「初期費用」は「礼金」と「敷金」の和を取るといったルール)など，様々なオントロジを用い関係を推論することにより，HTML といった非構造的なデータから，「面積」といったような key と「42 平米」といったような value が関連付けられた構造的なデータを取り出すことが可能となった。

そして通常各ページごと複数の物件がまとめて表になっているものを，Gryphon Broker で個々の情報としてユーザの興味と比較するために，今回は個々の物件ごとに別々の XML ファイルに一旦落とすことにした．

Gryphon Broker でマッチした情報はユーザへメールで配信されるが，今回は delivery monitor を使用し同じ Subject 毎にメールを合算しそれをユーザへ配信することにした．

3. 結論

本論文では標準化されつつあるオントロジ技術を用い，情報検索の場合の key となる語彙に対しこれを用いることで類似の key を含めて検索可能となり，また情報源の情報よりさらなる有益な情報を作成し，既存の情報配信システムよりも，より多くのかつ的確な情報を収集することが可能となった．

情報源がメタ情報を含まない非構造的な既存の HTML 形式で書かれたものからも、構造的なデータとしてそれを取り出し，ユーザの興味に合ったものを的確に配信することができ，さらに付加価値としてユーザにとってより便利な形に変形した情報を提供することが可能となった．

そして，オントロジを様々な対象領域ごとに用意し，これを差し替えることで，多様な情報源に対して複雑な作りこみをしなくても使用することが可能となった．

本論文で使用したように、WPS 上に配置したオントロジを用いる情報ポータルシステムと組み合わせた統合的なフレームワークを用いることで，ユーザはポータルページ上で的確に自分の興味を登録し，その興味にマッチした的確な情報を得ることが可能となる．

最後に今後の課題をあげると，本論文においては「情報が存在する場所」は自動で探すことができないため，手動で指定した情報源からしか対応させることができなかった．Web 上には様々な有益な情報がまだ数多く存在する．そういった情報を Monitor agent が自動で見つけ，これらの情報配信システムへ適応する事が今後の課題である．

Reference

[1]W3C SemanticWeb Activity (http://www.w3c.org/2001/sw)
[2] The DARPA agent Markup Language Homepage (http://www.daml.org)
[3] On-To-Knowledge Homepage (http://ontoknowledge.semanticweb.org)
[4] Distributed Messaging System (http://www.research.ibm.com/Gryphon)
[5] 松下望，槌谷一，SemanticWeb 上でオントロジを用いた検索 UI モデル，FIT (2002)
[6] News Industry Text Format (NITF) (http://www.nitf.org)