

文書中のキーワードに関する多種多様な情報を 収集・整理するシステム

～システムの概要と固有表現抽出技術,オントロジー技術～

松平
 洵上

正樹
 正睦

大沼 宏行
 森田 幸伯

上田 俊夫

近年、インターネットの発達により大量の情報が流通する中、利用者が必要な情報を適切に選択して取得することは困難になってきている。たとえば、報告書内の他社製品名に対してその製品情報や技術情報、評判情報を収集したり、ニュース等で注目されている技術について社内での取り組み、担当者を調べたいといったさまざまな要求に、検索エンジンが充分応えているとはいえず、大量の検索結果から必要な情報を利用者が探さなければならない。

この問題に対して、セマンティックWebやオントロジー等の技術が開発され、利用され始めている¹⁾。セマンティックWebは、Web上の文書に意味を付与することにより、コンピュータで処理できるようにするための技術で、「藤沢」が場所か人かを認識して問題解決をおこなうことができる。オントロジーは、元来は哲学用語で存在論という意味だが、情報処理分野では、概念（情報）と概念（情報）の意味的な関係を体系化したものを言う。

我々は、イントラネットやインターネット上の雑多な情報（非構造化情報）と、データベースやWebサービスのような構造化された情報（構造化情報）をオントロジーによって統合し、必要な情報を収集・整理して提供するシステム（以下、情報収集・整理サーバと呼ぶ）を開発している。本稿では、システムの概要を説明し、中心となる技術である固有表現抽出技術、ならびに、我々のオントロジー技術の特長について説明する。

情報収集・整理サーバの概要

我々が開発しているシステムは、ある文書内のキーワードに関連する多種多様な情報を収集し、意味的に統合・整理することを目的として、

- 文書テキスト中から人名、組織名、製品名、技術名等のキーワードを抽出して出力する機能
- キーワードに関連した情報をイントラネットおよびインターネットから収集する機能
- 収集した情報をキーワードの種類に応じて意味的な内容ごとに整理して出力する機能

を持っている。情報収集・整理サーバを、企業内情報ポータルに適用した場合の例を図1に示す。

まずシステムが、インターネット上のWebページ（ここではニュース）に対して製品名、組織名、人名等のキーワードを抽出して一覧を出力する（左上画面）。次に利用者が企業名（沖電気）を指定した場合は、企業情報とともに電話、地図情報、株価情報、取引履歴等のリンクボタンを表示する（中央画面）。これは、指定した企業（沖電気）に関連する多種多様な情報を、リンクボタンの意味する「箱」に整理していることになる。最後に地図情報ボタンを指定することにより、企業ホームページ内の地図を表示する（右下画面）。

中央画面において出力するボタン、すなわち整理する「箱」は、指定したキーワードの種類によって柔軟に変更し、必要な情報を容易に取得できるようにしている。



図1 情報収集・整理サーバ適用例

システム構成

情報収集・整理サーバのシステム構成を図2に示す。

システムは、指定されたテキスト中からキーワードを抽出するキーワード抽出部、指定されたキーワードに関連する情報を検索するリソース検索部およびWebサービス

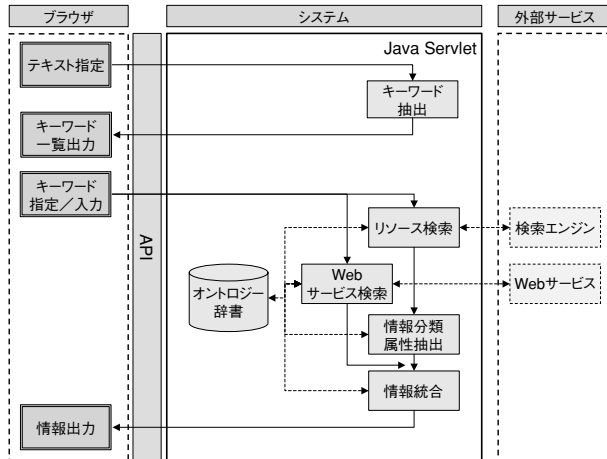


図2 システム構成

検索部、検索した情報を意味的に分類し、必要に応じて情報を抽出する情報分類・属性抽出部、意味的に同じ分類に属する情報を統合する情報統合部、および、情報と情報との意味的な関係を定義したオントロジー辞書から構成される。

キーワード抽出部は、後述する固有表現抽出技術をもとにテキスト中からキーワードを抽出し、一覧を出力する。その際、関連するキーワード、たとえば、隣接して出現する組織名と人名等は、関連づけている。抽出するキーワードの種類は、後述するタグとしてアプリケーションに応じて指定することが可能である。

リソース検索部は、外部の検索エンジンを利用して、必要な情報を検索する。その際、キーワードを検索エンジンに渡すだけでなく、ニュースを検索する場合に、ニュースサイトを指定したり、「ニュース」「プレスリリース」「報道資料」等の語句をキーワードに追加することにより、必要な情報を効率的に収集できるように工夫している。

書籍検索や旅行情報検索のようにWebサービスとして提供されている情報、ならびにデータベースの情報を、その要素を理解して必要な情報を収集するモジュールがWebサービス検索部である。これらの情報は構造化情報であり、提供される情報の要素と他の非構造化情報との意味的なマッピングを行っている。

情報分類・属性抽出部は、検索した情報をURLやタイトル、本文中の特定の語句等によるヒューリスティックな（経験則）ルールに基づいて意味的に分類する。情報を分類する「箱」は、指定されたキーワードの種類に応じて用意しているが、利用者が「箱」を追加定義あるいは削除することも可能である。この分類ルールは、情報の種類と対応づけてオントロジー辞書に格納している。

最後に、さまざまな情報源から収集して意味的に分類された情報を統合するのが、情報統合部である。同じ「箱」に分類された情報を、信頼性の最も高いものだけひとつ出力する、あるいは最新のものから順に出力する、同じ情報とみなせる場合は統合する等の処理でまとめることによって、ひとつのキーワードに関連する情報を俯瞰することが可能になる。

オントロジー辞書は、各部の処理に必要な情報やルールを定義した辞書で、処理に応じて参照される。詳細については後述する。

各部は、すべてJavaで記述しており、APIを介してJava Servletとして、キーワード抽出機能および情報収集・整理機能を提供している。

固有表現抽出技術： 文書からのキーワード抽出

キーワード抽出部の中心技術である固有表現抽出技術は、文書テキスト中の人名、組織名、製品名、技術名等の固有表現に意味的なタグを付与する技術である。たとえば、「沖電気 山田太郎」に対して、「<ORG>沖電気</ORG> <PERSON>山田太郎</PERSON>」のように組織を表すタグや人名を表すタグを付与する。キーワード抽出部は、アプリケーションによって指定されたタグが付与された語句をキーワードとして抽出する。

1999年に開催された第1回の情報検索、情報抽出に関するコンテストIREX (Information Retrieval and Extraction Exercise) ²⁾ の固有表現抽出課題では、以下の8つのタグが定義された。

- | | |
|------------|----------------|
| ●組織名、政府組織名 | <ORGANIZATION> |
| ●人名 | <PERSON> |
| ●地名 | <LOCATION> |
| ●固有物名 | <ARTIFACT> |
| ●日付表現 | <DATE> |
| ●時間表現 | <TIME> |
| ●金額表現 | <MONEY> |
| ●割合表現 | <PERCENT> |

我々はこれを独自に拡張し、

- | | |
|--------|-----------|
| ●サブ組織名 | <SUBORG> |
| ●姓 | <PS_L> |
| ●名 | <PS_F> |
| ●イベント | <EVT> |
| ●住所 | <ADDRESS> |

- 電話番号 <TEL>
- 電子メールアドレス <E_MAIL>
- 技術名 <WORD_Tech>
- 製品名 <PRODUCT>

等のタグを追加している（IREXのタグ名も一部変更している）。固有表現抽出の例を図3に示す。

固有表現抽出は、実用的な速度を達成するために、パターンマッチングと形態素解析による浅い解析処理を行っており、「〇〇株式会社」のように名詞の接頭や接尾の文字に着目して抽出する方式と、辞書を用いて抽出する方式を併用している。

アプリケーションによっては、現状のタグでは不十分なときや辞書を追加したい場合がある。たとえば、旅客交通のアプリケーションでは、鉄道名や駅名、空港名、道路名といったタグが必要になるかもしれない。これに対応するため、システムでは、接頭、接尾文字を利用したパターンマッチングのルールと辞書を外部ファイルとして定義し、タグの追加、編集等が可能なように、カスタマイズ性を高めている。

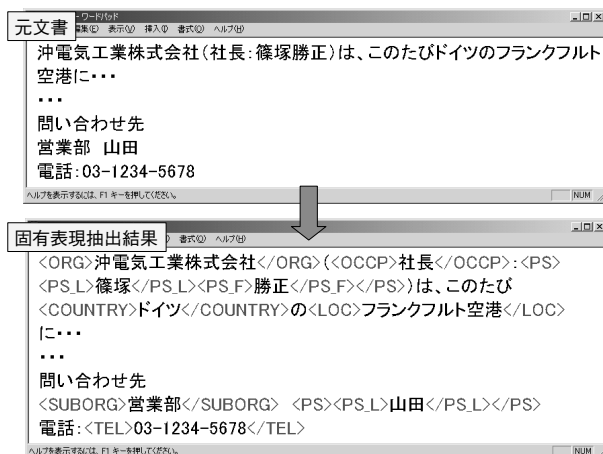


図3 固有表現抽出の例

オントロジー技術： 情報間の意味的な関係の定義手法

情報間の意味的な関係を定義した辞書がオントロジー辞書である。オントロジー技術は、それを矛盾なく、かつ効率的にどうやって表現すべきか、またどのように構築すべきかといった問題を取り扱う。

我々は、数千ページに及ぶイントラネット、インターネット上のWebページを分析し、人手でオントロジー辞書を構築した。オントロジー辞書は、情報と情報の意味

的な関係や情報の型、情報とそれを取得するためのルールの対応をRDF（Resource Description Framework）³⁾と呼ばれる3つ組のデータモデルで記述している。モデル化したオントロジー辞書の一部を図4に示す。

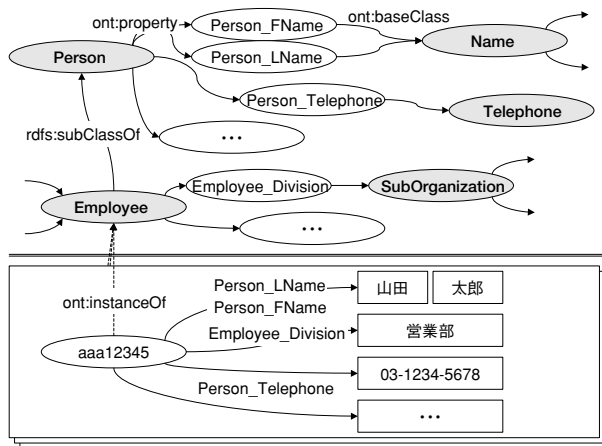


図4 オントロジーのRDFモデルの例

図4において、「Person」「Employee」「Telephone」等は情報のクラス、「Person_FName」「Person_Telephone」等は、クラスに関連する情報をプロパティとして示している。クラスは固有表現抽出のタグに相当し、プロパティは情報分類の「箱」に相当する。クラスとプロパティの関係はont:propertyで定義している。また、クラス間の関係として、「Employee」クラスは、「Person」クラスの下位関係、すなわち、より狭義の概念であることをrdfs:subClassOfで定義し、プロパティとクラスの関係として、「Employee_Division」プロパティは、値の型およびクラスが「SubOrganization」クラスであることをont:baseClassで定義している。

「aaa12345」は、インスタンスを識別するためにシステムが自動的に付与したIDであり、「Employee」クラスに対してont:instanceOfで定義されたひとつのインスタンス、すなわち、実体としてのある従業員を示している。データベースに情報がある場合は、必要に応じてそれを取得する。

また、図では省略しているが、各クラスおよびプロパティに対して、そのクラス、プロパティに属する情報を収集、分類するためのルール、およびデータベースから取得するためのWebサービスを対応づけている。この点が、我々が開発しているオントロジーの最大の特長である。すなわち、従業員に対して、所属や電話番号等の人事データ、担当製品、対外発表等の情報があり、人事データを従業員データベース、担当製品を製品データベースが

ら取得し、对外発表はインターネットから収集する方法がオントロジーに定義されており、ある従業員に関する多様な情報を整理して表示することが可能になる。

現在、ビジネス一般に必要なオントロジーとして、組織、人、場所、時間、文書、製品、技術、電子メール、電話等のクラス、および、それぞれのクラスについてプロパティを定義し、インスタンスとして企業名、地名、IT用語等を持っている。

また、オントロジーを用いたより高度な処理として、「山田さん問題」と呼ぶ個人の特定問題を解決するための手法を開発している。

企業内の文書では、「山田課長」のように名や従業員番号、メールアドレス等を伴わず、どの「山田」さんか判断できない場合が多い。これを「山田さん問題」と定義し、文書中に出現した「山田」さんが、個人を特定するための明示的な情報を伴わない場合に、文書内の情報から総合的に推論して特定しようというものである。我々の手法は、オントロジー上の人クラスあるいは従業員クラスに関連するプロパティを個人特定のための制約として利用し、文書内のキーワードとデータベースの従業員情報や他の文書から学習したインスタンスとのマッチングを行うことで、候補を絞り込む方式をとる。その際、キーワードの出現位置や種類、文書の種類に応じて重みづけを行うように工夫している。イントラネットの情報を利用した実験の結果、我々の手法は、すべての制約を同等に扱う手法と比較して、高い精度を得ている⁴⁾。

アプリケーション例

情報収集・整理サーバは、アプリケーション構築のための汎用的なコンポーネントと位置づけられ、オントロジー辞書を整備することによりさまざまなアプリケーションに利用可能である。

前述した企業内情報ポータルでの適用例では、IPコミュニケーションのシステムと連携させて、

- 顧客企業名から社内担当者を検索して連絡する
- 技術名から社内の専門家を探してアクセスする
- 小売業の経営層や管理部において店舗名から責任者に電話する

といったさまざまな利用が想定される。

また、コールセンターにおけるオペレータ支援、自治体ホームページのナビゲーション等、多種多様な情報を扱うアプリケーションへの適用を考えている。

特定のアプリケーション向けには、オントロジー辞書をカスタマイズする必要が生じる場合がある。前述したように、旅客交通のアプリケーションでは、鉄道名や駅

名、空港名、道路名といったクラスやプロパティが必要になるかもしれないし、医薬品分野では、薬品名や化学組成、効能、副作用等が必要であろう。このようなクラスやプロパティの追加、編集を行うために、オントロジー辞書のAPIを用意している。

今後の方向性

現在開発中のオントロジー辞書は、ビジネス一般で必要とされるものを対象としているが、まだ十分とは言えず、拡充を続けていく予定である。たとえば、会議室予約やプレゼンス管理に利用するために場所クラスの下位として部屋クラスを定義することや、より詳細な情報を収集するために製品クラスを細分化する等である。

また、アプリケーションの例として、社内の実データを利用した情報ポータルや、特定分野でのプロトタイプシステムを開発し、実証実験を通じて実運用の課題やノウハウを蓄積していく予定である。◆◆

参考文献

- 1) 三木, 松平, 大熊:「セマンティックメタデータ技術」, 沖テクニカルレビュー197号, pp.110-113, 2004年
- 2) 関根, 井佐原: IREX:「情報検索, 情報抽出コンテスト」, 情報自然言語処理, No.127, pp.109-116, 1998年
- 3) Manola, F., Miller, E.: RDF Primer (W3C Recommendation), <http://www.w3.org/TR/rdf-primer/>
- 4) 松平, 上田, 大沼, 淵上, 森田:「文書内の人名の個人特定に関する研究-「山田さん問題」の解決手法とその評価-」, 第3回情報科学技術フォーラム, 2004年

●筆者紹介

松平正樹: Masaki Matsudaira. 研究開発本部ヒューマンインタフェースラボトリ

大沼宏行: Hiroyuki Ohnuma. 研究開発本部ヒューマンインタフェースラボトリ

上田俊夫: Toshio Ueda. 研究開発本部ヒューマンインタフェースラボトリ

淵上正睦: Masachika Fuchigami. 研究開発本部ヒューマンインタフェースラボトリ

森田幸伯: Yukihiro Morita. 研究開発本部ヒューマンインタフェースラボトリ