人工知能学会研究会資料
SIG-SWO-A303-02

文書からのキーワード抽出と関連情報の収集

松平　正樹　　上田　俊夫　　大沼　宏行

渕上　正睦　　森田　幸伯

前回，キーワード入力に対して，オントロジーをもとに関連する情報を収集し，整理して出力するシステムの概要について報告した．本報告では，固有表現抽出技術を利用して文書からキーワードを抽出し，その意味属性に応じて情報を収集するように改良した内容について述べる．その際に起こるいくつかの問題について整理し，特に姓だけが出現する場合の個人特定の問題（「山田さん問題」と呼ぶ）について議論する．

キーワード：オントロジー，固有表現抽出，情報収集，Web サービス

Keyword Extraction from Documents and Information Collection about the Keyword

Masaki　MATSUDAIRA　　Toshio　UEDA　　Hiroyuki　OHNUMA

Masachika　FUCHIGAMI　　Yukihiro　MORITA

We reported the system which collect and arrange information about a keyword based on ontology. This paper describes keyword extraction from a document using named entity tagging technology, and information collection about the keyword. Some issues, especially person identify which we call “Yamada-san problem”, are presented.

Keywords: ontology, named entity tagging, information collection, web service

1.　はじめに

前回，情報を取得する際の問題として，検索エンジンでは必要な情報が多くの「ゴミ」に埋もれてしまうため，検索結果をひとつひとつ調査しなければならず，多大な手間がかかってしまう点，イントラネットに必要としている情報が存在する場合でも，ナレッジマネジメントシステムやコンテンツマネジメントシステムによってすべての情報が組織的に管理されていることは一般的には少なく，インターネット検索エンジンと同様の問題が起こる点を挙げ，キーワード入力に対して，オントロジーをもとに関連する情報を収集し，整理して出力するシステムの概要について報告した[松平 03]．技術情報をターゲットとして，図 1 に示すように，インターネットやイントラ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），そして，イントラネット上の仕様書や製品カタログといったフォーマットはある程度固定されているがそれぞれの項目が明示的に構造化されていない情報（半構造化情報）をオントロジーによって統合し，利用者に必要な情報を収集・抽出して提供するシステムである．

しかし，実際の利用シーンを考えると，例えば，

i.　会議報告書から他社製品名を抽出し，その製品に関する情報を収集する

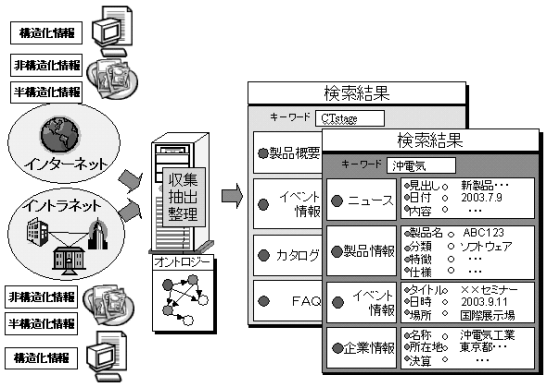


図 1　全体像

沖電気工業（株）　研究開発本部
Oki Electric Industry, Co., Ltd. Research and Development Division
E-mail: matsudaira564@oki.com

ii. メールから人名を抽出し，電話／メールで連絡をとる
iii. 会議開催案内メールから日付・時間を抽出し，スケジュールが空いているか確認する

というように，文書閲覧時に，その文脈で情報を収集したいという要求がある．

我々は，このようなアプリケーションを想定し，固有表現抽出技術を利用して文書からキーワードを抽出し，その意味属性に応じて情報を収集するようにシステムを改良した．

本報告では，まず 2 章でシステムの改良部分を中心に，固有表現技術の概要を述べる．3 章では，その際に起こるいくつかの問題について整理し，特に姓だけが出現する場合の個人特定の問題（「山田さん問題」と呼ぶ）について説明する．問題の解決方式を 4 章で議論する．

2. システム概要

改良したシステムでは，Web ページを指定すると，その HTML 文書中から人名，組織名，製品名，技術名等のキーワードとその意味属性を抽出し，出力する．利用者がキーワードのひとつを指定することにより，前回報告したシステムがその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．システムの構成を図 2 に示す．

以下で，キーワード抽出処理とそのコア技術である固有表現抽出技術，および情報収集処理について述べる．

2.1. 固有表現抽出技術

まず，キーワード抽出処理のコア技術となる固有表現抽出技術について説明する．

固有表現抽出技術は，文書中の固有表現に意味的なタグを付与する技術である[関根 98][大沼 03]．例えば，文書中の「沖電気　山田太郎」に対して，「<ORG>沖電気</ORG>　<PERSON>山田太郎</PERSON>のように組織タグや人名タグを付与する．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）の固有表現抽出（NE）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>

その後，関根は独自に拡張および階層化をおこなっている[関根 03]が，我々も独自に，

- 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>
- URL <URL>
- 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)．図 3 に固有表現抽出の例を示す．

2.2. キーワード抽出処理

キーワード抽出処理は，固有表現技術を利用してタグづけされた文書から，指定した意味タグの要素をキーワードとして抽出する処理である．現在のシステムでは，組織名，サブ組織名，人名（姓，名），製品名，技術名を出力しており，各タグをオントロジーのクラス定義および属性項目定義の URI（例えば，ont:Person および ont:Person_Name）にマッピングしている．情報収集処理には，この URI を渡すことにより，クラスあるいは属性項目を特定し，情報収集処理をおこなう．

ここでひとつの問題がある．図 3 の固有表現抽出

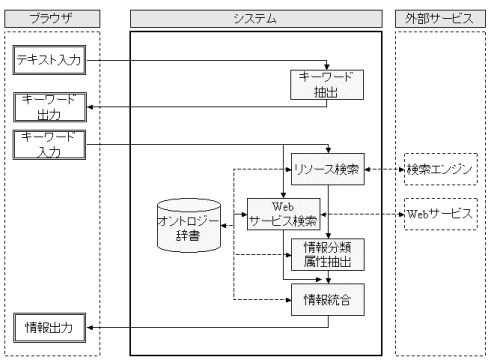


図 2　システム構成図

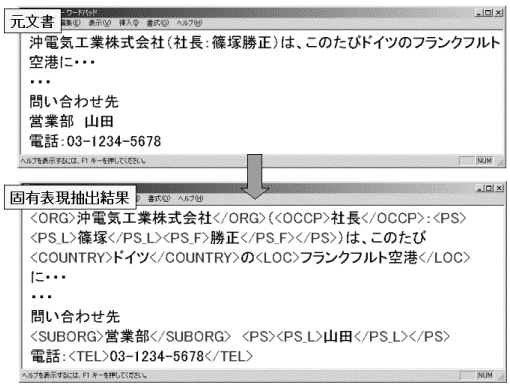


図 3　固有表現抽出例

結果から上記抽出タグにしたがってキーワード抽出をおこなうと，抽出される情報は，

沖電気工業株式会社（組織名）
営業部（サブ組織名）
篠塚 勝正（人名（姓，名））
山田（人名（姓））

となり，元の文書に記述されている電話番号やメールアドレス等の情報や，「営業部」と「山田」の関係が失われてしまう．この問題に対処するため，キーワードに関連する情報を補助情報として抽出することとした．すなわち，人名キーワード「山田」に対して，サブ組織名「営業部」，電話番号「03-1234-5678」を補助情報として関連づけて出力する．これらの情報をもとに，次の情報収集処理のステップで，姓だけが出現する個人の特定や条件を付与した検索処理等をおこなう．個人特定の問題（「山田さん問題」）については，後述する．

2.3. 情報収集処理

情報収集処理は，前回報告した処理と同様である．オントロジー辞書，および情報抽出・属性抽出，情報統合の各内部モジュール，検索エンジン，Web サービス等の外部サービスを利用してイントラネットおよびインターネットから情報を収集し，また Web サービスから情報を取得し，利用者に必要な情報を属性項目ごとに情報を抽出・整理して出力する．インターネット上の情報は Google 検索エンジン [Google]を利用し，書籍情報として Amazon の Web サービス [Amazon] を利用している．

前回の報告から異なる部分は，キーワードに対して，オントロジーのクラス定義あるいは属性項目定義の URI を付与し，関連情報を補助情報として渡している点である．出力例を図 4 に示す．

3. 山田さん問題

個人を特定するためには，どのような情報が必要であろうか？

Semantic Web の世界では，リソースの URI によって個人を識別することができるが，まず URI あるいは URI と 1 対 1 に対応する情報を取得する必要がある．企業内の従業員の場合は，氏名や従業員番号，メールアドレス等で特定することができるかもしれない．FOAF（Friend of a Friend）プロジェクト [FOAF]では，ある人の友人を特定するためにメールアドレスを利用している．しかし，一般的に既存の大量の文書を対象にしたいという要求が強く，文書内の個人名は姓だけで，メールアドレスや従業員番号が記述されていないケースも多い．そのような場合，記述されている姓と周辺の情報から個人を特定しなければならない．実際，単に「山田」で従業員データベースを検索すると 100 件以上結果が見つかり，何らかの制約が必要である．この場合，所属や役職等の制約により，個人を特定，あるいは，候補を数人に絞ることができる．今回は，従業員データベースや顧客データベースがあることを想定し，姓から個人名の候補を絞るという問題を「山田さん問題」と定義する．「山田さん問題」の特徴的な例を図 5 に示す．以下で実際の例を分析し，解決方式について議論する．

3.1. 出現パターン分析

「山田さん問題」に対して，イントラネットから日本人の姓として多い「山田」「佐藤」「鈴木」「高橋」「田中」を含む文書それぞれ 100 件を検索し，出現パターンを分析した．

まず，それぞれの姓だけ（山田（太）のように姓と名の一部で出現するものを含む）が出現する文書の件数は，以下の通りである．

- 山田： 28 件 / 100
- 佐藤： 42 件 / 100
- 鈴木： 29 件 / 100
- 高橋： 32 件 / 100
- 田中： 42 件 / 100

- 合計： 173 件 / 500 （34.6%）

結果から，約 1/3 の文書で「山田さん問題」を含ん

図 4 システム出力例

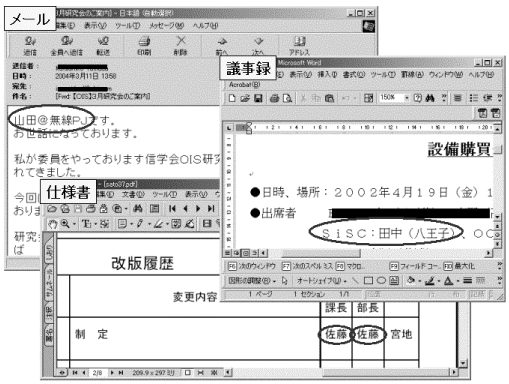


図 5 山田さん問題の例

でいることがわかる．文書属性ごとの出現パターンを以下に示す．

議事録
出席者および発言者として記述される．「営業部 山田課長」のように，組織名あるいは役職名を伴うものも多い．

仕様書
改版履歴の記入者，承認者として記述される．組織名を伴って記述されるか，伴わない場合は仕様書の表紙ページに組織名がある．

管理表
担当製品，担当顧客等の情報が対応づけられている．組織名や電話番号，メールアドレスが記述されている場合もある．

座席表
「03-1234-5678　山田」のように，電話番号との対応づけがされている．ただし，ファイル形式によっては電話番号と姓が別々のテキスト枠になっており，対応づけが困難な場合もある．また，組織名との対応づけは一般的には困難である．

ホームページ
「問い合わせ先：　山田（taro.yamada@oki.com）」のように，連絡先としてメールアドレスが記述される場合が多い．

論文
参考文献に論文のタイトル，発表先等を伴って記述される．

メール
「山田＠営業部です」「開発部　山田様」のように，差出人の紹介，宛名，あるいは文中に記述される．多くの場合，差出人，宛先，同報先のアドレスに対応するメールアドレスが存在する．

以上の分析から，「山田さん問題」は，組織名，役職名，電話番号，担当製品，担当顧客等の人（あるいは従業員）に関するオントロジーの属性項目の情報を総合的に制約として利用することで解決できそうである．

3.2. その他の問題

固有表現抽出技術を利用してキーワードとその意味属性を抽出した後，キーワードのひとつを指定することにより，そのキーワードに関連した情報を収集するというプロセスは，「山田さん問題」以外にもいくつかの問題を含んでいる．

例えば，固有表現抽出技術で抽出できないキーワードがある．これは，「ゴーン」「CTstage」「沖ソフトウェア」といった登録されていない人名や製品名，企業名が，「氏」「さん」「株式会社」など意味属性を示すキーワードを伴わずに文書中に出現した場合に起こる．辞書の整備や，意味属性を伴って出現したパターンからの学習による解決が考えられるが，詳細については今後報告する予定である．

また，「松本」「山口」など人名と地名の属性を誤りやすい単語，「林」など一般名詞になる人名の問題も，本稿では取り上げない．

4. 問題の解決方式

この章では，上述した「山田さん問題」の解決方式について議論する．

4.1. 人に関するオントロジー

出現パターンの分析から，「山田さん問題」は，人に関するオントロジーの属性項目の情報を総合的に制約として利用することで解決できそうである．まず，オントロジーとして Person（人）クラスの定義の一部を図 6 の上部に示す．

- クラスは複数の属性項目を有する
  Person（人）クラス
  ＜属性項目＞
  Person_Name（氏名）属性,
  Person_Division（所属）属性,
  Person_Telephone（電話）属性,
  . . .
- 一部の属性は，階層化される
  Person_Name（氏名）属性
  ＜下位属性＞
  Person_Lname（姓）属性,
  Person_Fname（名）属性
- 一部の属性は，ベースクラスを有する
  Person_Division（所属）属性
  ＜ベースクラス＞
  SubOrganization（サブ組織）クラス

ベースクラスは，属性項目に入る値の型である．

従業員データベースや顧客データベースは，レコードをインスタンスとみなして操作することを考える．図 6 の下部はその例であり，営業部の山田太郎

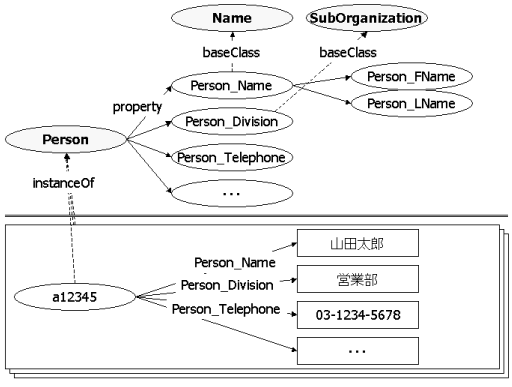


図 6　人に関するオントロジー

は,「山田太郎」という Person_Name（氏名）属性,「営業部」という Person_Division（所属）属性,「03-1234-5678」という Person_Telephone（電話）属性を持つPersonクラスのインスタンスということを示している.（システムでは，属性値として文字列を記述している場合と，Web サービスからの取得方法を指定している場合がある）

オントロジーは，RDF 形式で記述し，Sesame をベースに構築したリポジトリで管理している.インスタンスの「a12345」は Sesame が自動的に付与する ID である.

4.2. インスタンス候補の絞込み

文書に出現したキーワード（例えば,「山田」），あるいは固有表現抽出処理で抽出した語（例えば,「社長」「フランクフルト」「営業部」）の中から，意味タグが Person（人）クラスの属性項目になるもの，あるいは，その語の意味タグが Person（人）クラスの属性項目のベースクラスと同じものを，Person（人）クラスのインスタンスを絞り込むための制約とする.図 3 の元文書から，Person（人）クラスに対して抽出する制約を図 7(a)に示す．制約として，意味タグが Person_Post（役職）である「社長」，Organization_Name（企業名）である「沖電気工業株式会社」，SubOrganization_Name （サブ組織名）である「営業部」，Telephone_Number（電話番号）である「03-1234-5678」を利用する．その際，キーワードとの文書内での距離，および各属性項目についてのヒューリスティックなルールにより，重みづけをおこなう．そのため，キーワードおよび制約となる語の元文書内の行番号および行内位置を抽出している．インスタンスの絞込みのアルゴリズムは以下の通りである.

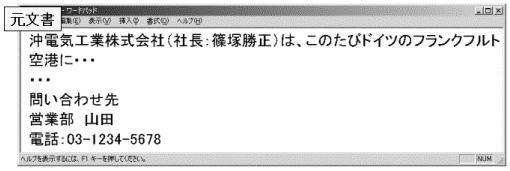
```
Person_LName あるいは Person_Name
にマッチするインスタンスを検索する

if( 検索したインスタンスの候補が 1 件 ){
   インスタンスを出力して終了する
}

for( 各インスタンス候補 ){
   for( 各制約 ){
      インスタンスと制約のマッチングを
      おこない，マッチした場合はこの重
      みづけによりインスタンスに得点を
      与える
   }
}

インスタンスの候補を得点の高い順にソ
ートして出力する
```

この例では，SubOrganization_Name（サブ組織名）「営業部」がキーワード「山田」と同一行にあり，また，人名とサブ組織名とのヒューリスティックルールから Person_Division（所属）として最も強い制約，すなわち重みづけ（得点）の高い制約になる．同様の手続により，Telephone_Number（電話番号）が次に強い制約として作用し，Person_Post（役職）「社長」および Organization_Name（組織名）は弱い制約となる．属性項目と制約の関係を図 7(b)，全体の制約を図 7(c)に示す. これらすべての制約によっ



「山田」への制約:

キーワード:	クラス・属性項目	line	position
山田	Person_LName	8	7
制約:		line	pos
沖電気工業株式会社	Organization_Name	1	0
社長	Person_Post	1	20
営業部	SubOrganization_Name	8	0
03-1234-5678	Telephone_Number	9	6

図 7(a)　元文書から抽出した制約

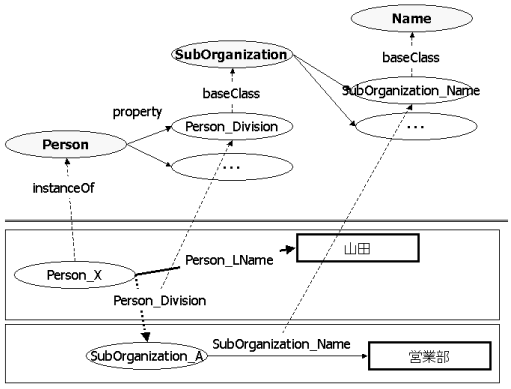


図 7(b)　属性項目と制約の関係

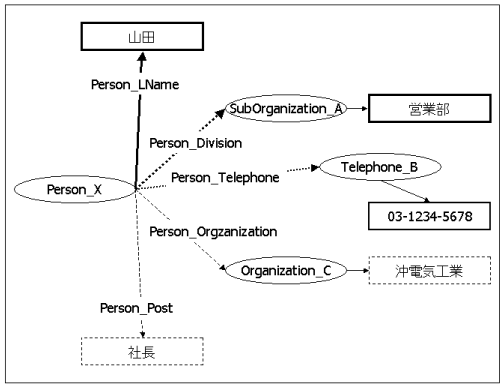


図 7(c)　全体の制約

てインスタンスの総合得点を計算し，候補を絞込む．

4.3. 議論

インスタンスとのマッチングによる制約による解決方式を述べたが，解決できない場合がいくつかある．インスタンスに存在しない個人の特定，十分な制約が得られない場合である．

前者は，従業員や顧客データベースに登録された個人以外の人名であり，例えば，出現パターン分析の際に述べた論文の参考文献，メール等に出現する人名がある．論文の参考文献での人名は，科学技術振興機構（JST）の提供する JST Online Infomation System（JOIS）情報検索サービス[JOIS]や，各学会の公開データベース，Web ページ等を利用することが考えられるが，そのためには，それらの情報を内部オントロジーへの変換をおこなう必要がある．1 対 1 にマッピングできる情報もあるが，例えば，JOIS が提供している論文情報の項目のひとつである＜著者名および所属機関名 AU＞は， 本システムで利用している Person_Name, Person_Organization 属性を含んでおり，それぞれの要素に分割しなければならないものもある．このような情報サービスは，今後 Web サービス化が進むと予想され，詳細な属性が容易に取得できるようになることが望まれる．一方，メール内に出現する人名は，送信者，受信者間での文脈に依存して特定しなければならないことがある．すなわち，送信者，受信者に関連する過去のメールや文書から制約を抽出しなければならず，文脈をどのように定義するか，文脈のどの範囲を制約として利用するかといった課題を解決しなければならないだろう．

後者，すなわち，十分な制約が得られない場合は，文書に制約となる情報が出現しない場合，および制約は出現するのだがインスタンスの属性に十分な情報が存在しない場合に分類できる．

文書に制約となる情報が出現しない場合は，情報は文脈に依存しており，先述した文脈の定義，制約としての利用についての課題を解決しなければならない．

インスタンスの属性に十分な情報が存在しない場合とは，例えば，文書に出現する製品名や顧客名，技術名が，従業員あるいは顧客のインスタンスとして記述されていないため，制約として抽出できても絞込みができない場合である．この場合，他の文書から抽出した個人を特定できる人名と，製品名や顧客名，技術名の関係をインスタンスとして学習することが考えられる．例えば，開発計画書に製品名と担当部署，担当者の関係が記述されていれば，それらの関係をインスタンスとして学習し，議事録に顧客との折衝記録があり，顧客名と担当者名，所属の関係が抽出できれば，その関係も学習する．ただし，自動的に学習するのは悪影響も大きく，TF・IDF 法や SVM のような統計的な手法，あるいは人間を介在して登録する方式を検討する必要があるだろう．

5. まとめ

前回報告した，キーワード入力に対してオントロジーをもとに関連する情報を収集し，整理して出力するシステムを改良した内容，すなわち，固有表現抽出技術を利用して文書からキーワードを抽出し，その意味属性に応じて情報を収集するようにしたシステムについて述べた．その際に起こるいくつかの問題について整理し，特に姓だけが出現する場合の個人特定の問題を「山田さん問題」と定義した．その出現パターンを分析し，文書内に出現するキーワードを制約としてオントロジーを用いて解決する方式について議論した．今回は解決方式の指針を示しただけであり，今後は定量的評価をおこなう予定である．

参考文献

[Amazon] Amazon.co.jp, Amazon Web サービス, http://www.amazon.co.jp/exec/obidos/subst/associates/join/webservices.html/

[Chris03] Chris, H., Lin, L., Matt, P., Eric, S., PRONTO: PoRtal based on ONTOlogy, http://www.cs.uga.edu/~ch/GlobalInfoSys/, 2003

[FOAF] FOAF Project, http://www.foaf-project.org/

[Google] Google, Google Web APIs, http://www.google.com/apis/

[JOIS] JST Online Infomation System (JOIS), 科学技術振興機構, http://pr.jst.go.jp/db/jois/

[Niles 01] Niles, I. and Pease, A., Origins of The IEEE Standard Upper Ontology, Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-01), 2001

[Sesame] Sesame 1.0-pre4, openRDF.org, http://www.openrdf.org/

[TAP] TAP Semantic Search, http://tap.stanford.edu/

[大沼 03] 大沼, 松平, 渕上, 森田, Web コンテンツの分析に基づくオントロジ構築および属性抽出の試み, 情報処理学会 情報学基礎研究報告 No.72／自然言語処理研究報告 No.157, 2003

[関根 98] 関根, 井佐原, IREX：情報検索、情報抽出コンテスト, 情報処理学会 情報学基礎研究報告 No.51／自然言語処理研究報告 No.127, 1998

[関根 03] 関根, 関根の拡張固有表現階層, http://apple.cs.nyu.edu/neh/

[武田 01] 武田, 人工知能におけるオントロジーとその応用, 2001

[松平 03] 松平, 上田, 大沼, 森田, Web コンテンツの分析に基づくオントロジー構築および情報整理の試み, 人工知能学会, 第5回セマンティックウェブとオントロジー研究会, 2003