社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－ＦＩ－90 (4)
2008－ＤＤ－65 (4)
2008／3／28

オントロジーを用いたニュース理解支援方式

吉田 慶章† 柿崎 淑郎‡ 辻 秀一††

†東海大学大学院 工学研究科
259-1292 平塚市北金目 1117 番地

‡東海大学連合大学院 理工学研究科
259-1292 平塚市北金目 1117 番地

††東海大学 情報理工学部
259-1292 平塚市北金目 1117 番地

あらまし ニュース記事を理解するにはユーザの知識が必要になるため，ユーザの記事理解を支援するサービスが行われている．しかしサービスにおいて必ずしもユーザの理解を促進させる情報が提供されているとは言えない．そこで本稿ではオントロジーを用いて記事に適した関連情報を提供する理解支援方式を提案する．オントロジーは記事から知識を抽出し構築する．そしてプロトタイプシステムを試作し，提案手法と関連手法を比較することで有効性を議論する．

A New Understanding Support for News Using Ontology

Yoshiaki YOSHIDA† Yoshio KAKIZAKI‡ Hidekazu TSUJI††

†Graduate School of Engineering
Tokai University
1117 Kitakaname, Hiratsuka,
Kanagawa, 259-1292 Japan

‡Graduate School of Science and Technology
Tokai University Unified Graduate Shool
1117 Kitakaname, Hiratsuka,
Kanagawa, 259-1292 Japan

††School of Information Science and Technology
Tokai University
1117 Kitakaname, Hiratsuka,
Kanagawa, 259-1292 Japan

Abstract There is service which supports for user understanding of the articles because understanding of the articles needs user's knowledge. However it is difficult for user to understand of the articles by using support service. In this paper, we propose a new understanding support method which provides knowledge applied of articles using Ontology. We construct Ontology with extracted knowledge from the articles. We implement prototype system, and discuss effectiveness by comparing our method with related methods.

1 はじめに

現在ユーザが積極的に情報を発信していく時代，言わばWeb2.0と言われている．Web2.0の問題点として多種多様な情報が氾濫していることが挙げられる．Web2.0では，膨大な情報の中から注目するべき情報をいかに効率良く得るか，そして情報の理解をいかに効率良く深めるかが重要である．前者は非検索という切り口からニュース情報を可視化し，タグクラウドインタフェースを用いて視覚的に提供する方式を提案した[1]．そこで本稿では後者に着目し，ユーザのニュース理解を支援する方式を提案する．

一般にニュースは速報性が重視されているた

めに，記事が古くなればなるほど価値が失われていくという特徴がある．多くのウェブサイトにおいて公開から二週間から一ヶ月程経過するとパーマリンクがなくなり記事として公開されなくなってしまう．

そこでニュース記事の価値が単に失われていくのを見過ごすのではなく，日々の出来事や時世の動きを知識として体系化しニュースを読む際の理解支援として再利用することが有効だと考えた．

本稿では情報を体系化するためにセマンティックウェブ技術の基盤であるオントロジーを用いる．オントロジーを用いることで情報の意味や概念，関係を表すことができる．文献 [2] では Web3.0 を“意味を表現し知識を連結し，これらをより自分にとって意味の深く便利で，そして楽しいインターネット体験にするために使うこと”と定義している．オントロジーを用いたアプリケーションを構築していくことで，コンピュータが情報の意味や知識を理解するセマンティックウェブの実現がより現実的になることが期待される．

2 従来方式

2.1 従来サービス

記事をはじめ文書を閲覧する際のユーザの理解度は総合的な知識に依存するため，文中に知らない用語が出てくることが想定される．その場合，多くはそのまま読み飛ばしてしまう，もしくはその用語を二次的に検索をして理解しようと試みるであろう．このようなユーザの理解を支援するために，文中のキーワードにリンクが貼られ，そこから関連情報を提供するサービスが行われている．Wikipedia やはてなダイアリーのオートリンクがその一例である．この例ではそれぞれのキーワードページに遷移する．

しかしこれらのサービスは二次的な検索であることには変わりがなくそこからまた情報を探さなければならない．さらにそこから得られる情報というのは閲覧している記事やエントリーに関連するものだけではなく，そのキーワード全てにおける関連情報である点でユーザに効率的な関連情報の提供が行われているとは思えない．

さらに同じトピックの過去の報道や動向を知らない場合に記事の概要を捉えられないという問題点が存在する．この問題点を解決しようと，MSN 産経ニュース[1]のサービスでは一般のウェブ検索だけでなく記事内検索と画像検索ができるようになっている．記事内検索では，キーワードに関連する他ニュースを比較して読むことができるため理解が促進されると考えられる．しかし必ずしもその記事を理解するのに必要な情報が提供されているわけではなく，同時期の記事の同じキーワードからは同じ情報が提供されているという点で問題である．

さらに，はじめに記した通りニュース情報は一定期間経過すると情報が公開されなくなってしまうため，過去のニュースは当然検索対象外となってしまう．

2.2 関連研究

北山らの研究 [3] では映像ニュースとテキストニュースそれぞれのコンテンツ構成順序の特徴に基づいた比較ニュース検索の質問生成を行うことで，記事の理解を促進させるニュースコンテンツを提供している．

森らの研究 [4] では個人情報を公開するオントロジーである FOAF を用いて，研究者の情報をキーワードとして自動的に抽出する手法を提案している．抽出されている情報は所属や研究分野に関するキーワードであり，人物を表す上で有用であると考えられる．

奥田らの研究 [5] では新聞記事と Blog から価格など時間と共に変動する動向情報を抽出している．本稿で扱う動向情報は数値情報だけでなくトピックに対する人物の発言にも着目している点で異なる．

綾らの研究 [6] ではセマンティックオーサリングにより作成されたコンテンツからの文章生成を目的として，修辞構造のアノテーションに基づいた新聞記事の要約生成を行っている．

[1] http://sankei.jp.msn.com/

3 ニュース理解支援方式

本稿ではニュース記事から知識を抽出し，ニュースオントロジーを構築した．このオントロジーを用いてユーザのニュース理解を支援する方式を提案する．

キーワードから提供される情報が必ずしもその記事の理解が促進されるものではないという問題点に対して，それぞれの情報を関連しているトピックと紐付けする構成でオントロジーを構築することで解決する．その結果，ユーザに提供される情報は閲覧記事に合ったものに限定される．情報を関連する項目と関連付けることができる点がオントロジーを用いる利点の一つである．

そして関連情報とは記事中に表れるキーワードにおいて，人物名であれば役職歴やトピックに向けた発言や取り組みを指し，政策や会議などのキーワードであれば，決定事項やその用語の説明を指す．

本稿ではニュースを理解するためには，動向情報として時系列に沿った人物の発言の変化，キーワードであればその説明や内容が提供されることが必要であると考えてこれらの関連情報を用いることとした．

図 1 に本稿で提案するニュース理解支援方式の構成図を示す．ユーザが本稿で試作したプロトタイプシステムにアクセスした際に，記事データベースから記事情報が抽出され，オントロジーからはその記事トピックに関連した情報が出力される．サーバにて記事中のキーワードに対して自動的に関連情報へのリンク処理を行い，関連情報を付与した後，ユーザ出力される．この結果，ユーザは関連情報を付与された記事を閲覧することで，関連情報を取得し，理解の促進に用いることができる．

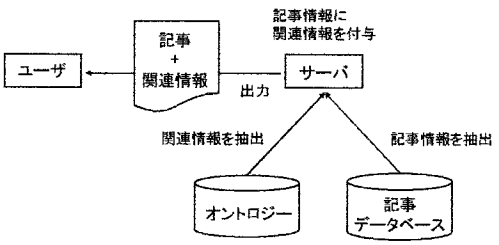

図 1: 構成図

3.1 ニュースオントロジー

今回構築したニュースオントロジーは日々のニュース動向を追い記事に関連する用語や関連のある用語を体系化したものである．今回は対象とするニュースジャンルを政治に限定する．

ニュースオントロジーは OWL 言語を用い，protege[2]により手動で構築している．

図 2 に現総理大臣である福田康夫と現内閣官房長官である町村信考を取り巻く周辺のオントロジーを示す．

福田康夫と町村信考は people クラスのインスタンスである．内閣府は内閣総理大臣と内閣官房長官をインスタンスとして保持している．現在就任していることを表す `ex:Current-Assumption` プロパティで福田康夫と内閣総理大臣，町村信考と内閣官房長官を連結している．

人物と発言は `ex:Saying` プロパティで連結されている．それぞれの発言は発言クラスのインスタンスであり，リテラルとして発言された日時を保持している．

そして関連していることを表す `ex:Related` プロパティを用いて，それぞれの発言がどのトピックに関連したものであるかを紐付けする．図 2 よりそれぞれの発言が年金問題やギョーザ中毒問題と関連していることがわかる．

またトピックと紐付けされていない，就任日や第何代の就任なのかなどの関連情報はどのニュースの関連情報となり得る基礎的な知識であると考えている．

図 2 において二重線で囲まれた四角はリテラルである．これらをつなぐプロパティ `ex:-AssumptionDate`（就任日）や `ex:SayingDate`（発言日）は DatatypeProperty 型である．図 3 は図 2 を表す OWL ファイルの一部である．前述したオントロジーの構成を記述している．

[2]http://protege.stanford.edu/

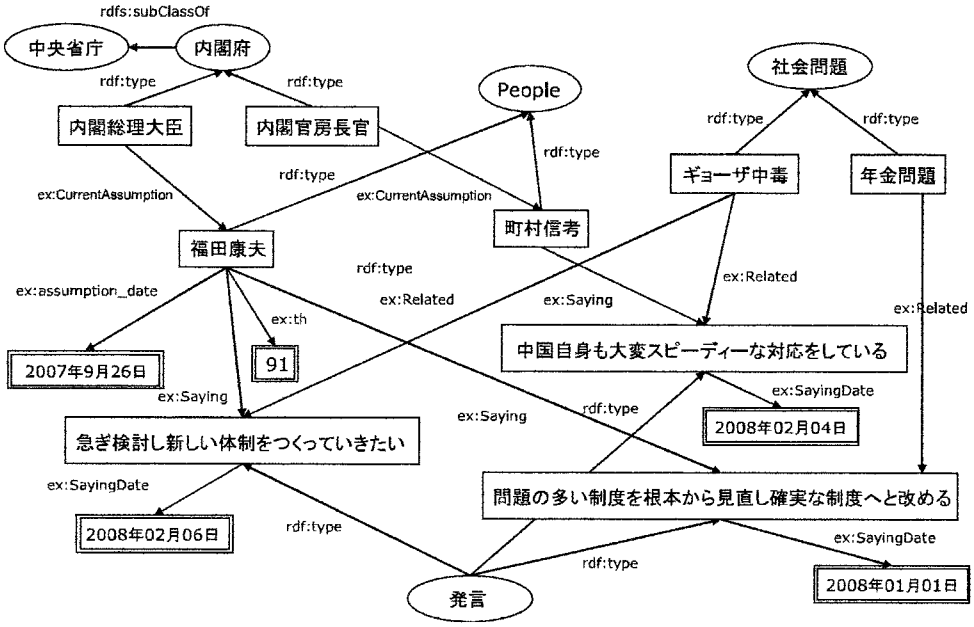

図 2: ニュースオントロジー例

3.2 関連情報抽出

本稿では，関連情報抽出に SPARQL (SPARQL Protocol and RDF Query Language)[3]を用いる．SPARQL は 2008 年 1 月に W3C 勧告として発表された RDF のためのクエリ言語である．

例としてオントロジーから福田康夫と年金問題のニュースに関する関連情報（ここでは発言）を取得する SPARQL クエリを図 4 に示す．

まず，人が発言したことを示す `ex:Saying` プロパティを用いて福田康夫の発言を抽出している．その発言は`?say` にバインドされる．その `?say` に対して発言日時を表す `ex:SayingDate` プロパティで検索することで，発言日時を得ることができる．また発言`?say` がどんな問題と関連しているのかを抽出するために `ex:Related` プロパティで検索をしている．ここで FILTER を用いることで，抽出された福田康夫の発言の中から，閲覧している記事にマッチした問題のみの情報を抽出することができる．ここでは年金問題に関連のあるもののみを抽出した．

またこの例では年金問題であるが，閲覧している記事がどんなトピックに属するかは今回は手動で判別を行った．

クエリ結果を ORDER BY を用いて発言日でソートを行っている．問題に対する発言の変化が明らかになり，記事の理解が促進されると考える．

そして図 5 が図 4 のクエリの結果例である．福田康夫の年金問題に関わる発言のみが抽出されている．2004 年の発言は内閣官房長官時代のものであり，国民年金保険料を払っているか否かの公表に対する発言である．他クエリを投げることで，役職歴やスローガンを抽出することができる．

4 実装と評価

4.1 プロトタイプシステム

今回 PHP を用いてプロトタイプを実装した．オントロジーからの関連情報抽出には RAP-API for PHP V0.9.5[4]を利用した．

[3]http://www.w3.org/TR/rdf-sparql-query/

[4]http://www4.wiwiss.fu-berlin.de/bizer/rdfapi/

```
<People rdf:ID="福田康夫">
 <Saying>
  <発言 rdf:ID=
      "急ぎ検討し新しい体制をつくっていきたい">
   <Related rdf:resource="#ギョーザ中毒"/>
    <SayingDate rdf:datatype="http://www.w3.org
     /2001/XMLSchema#int">20080206</SayingDate>
  </発言>
 </Saying>
 <Saying>
  <発言 rdf:ID="問題の多い制度を
    根本から見直し確実な制度へと改める">
   <Related rdf:resource="#年金問題"/>
    <SayingDate rdf:datatype="http://www.w3.org
     /2001/XMLSchema#int">20080101</SayingDate>
  </発言>
 </Saying>
</People>

<内閣府 rdf:ID="内閣官房長官">
 <CurrentAssumption>
  <People rdf:ID="町村信孝">
   <Saying>
    <発言 rdf:ID="中国自身も大変スピーディーな
      対応をしているということは言える">
     <Related rdf:resource="#ギョーザ中毒"/>
      <SayingDate rdf:datatype="http://www.w3.org
      /2001/XMLSchema#int">20080204</SayingDate>
    </発言>
   </Saying>
  </People>
 </CurrentAssumption>
</内閣府>
```

図 3: オントロジーファイル例

図 6 はプロトタイプシステムの画面例である．左側はニュース見出し一覧であり，選択したニュース記事が右側表示されるようになっている．記事にマッチした関連情報はニュース記事読み出し時にオートリンクされたキーワードにマウスを乗せることで提供される．今回は JavaScript を用いて記事の下にポップアップされるようなインタフェースで実現した．

4.2 評価

現時点ではまだ理解促進に対する網羅性評価が行えてないため，本稿では関連研究との比較とオントロジーの構築的評価に留める．

```
SELECT ?say ?s_date ?s_relate
WHERE{
    ex:福田康夫 ex:Saying ?say .
    ?say ex:SayingDate ?s_date .
    ?say ex:Related ?s_relate .
    FILTER (regex(str(?s_relate),"年金問題","i")) .
}
ORDER BY DESC(ex:SayingDate)
LIMIT 20
```

図 4: SPARQL クエリ例

?say	?s_date	?s_relate
来年３月までの名寄せの完了	20070103	年金問題
問題の多い制度を根本から見直し	20070101	年金問題
個人情報でそういうものを	20040428	年金問題
開示すべきではない		

図 5: SPARQL クエリ結果例

4.2.1 関連研究との比較

まず北山らの研究 [3] との比較を行う．北山らは閲覧中の記事の理解を促進させるニュース記事を提供するという試みをしており本稿と目的は同じである．しかし本稿で提供する情報は記事コンテンツではなく，記事コンテンツから抽出したニュースの動向や知識であるために手法が異なる．記事理解を促進させるコンテンツが提示されたとしても，そのコンテンツを二次的に読んで理解するというステップを要するため，芋づる式に読むコンテンツが増えていく可能性がある．それに対して本稿は知識を直接提供するために，その場で理解の促進に役立てることができる点で有用性がある．

FOAF を用いることで研究者の情報をキーワードとして自動的に抽出している森らの研究 [4] と比較すると，人物の所属や役職などのキーワードを保持できる点で同じである．しかし本稿では抽出を手動で行ったが，人物の発言も対象の範囲としている点でニュースを扱う上では有用性がある．

綾らの研究 [6] をはじめ，文書にアノテーションを付与する研究では，処理対象となる文書に

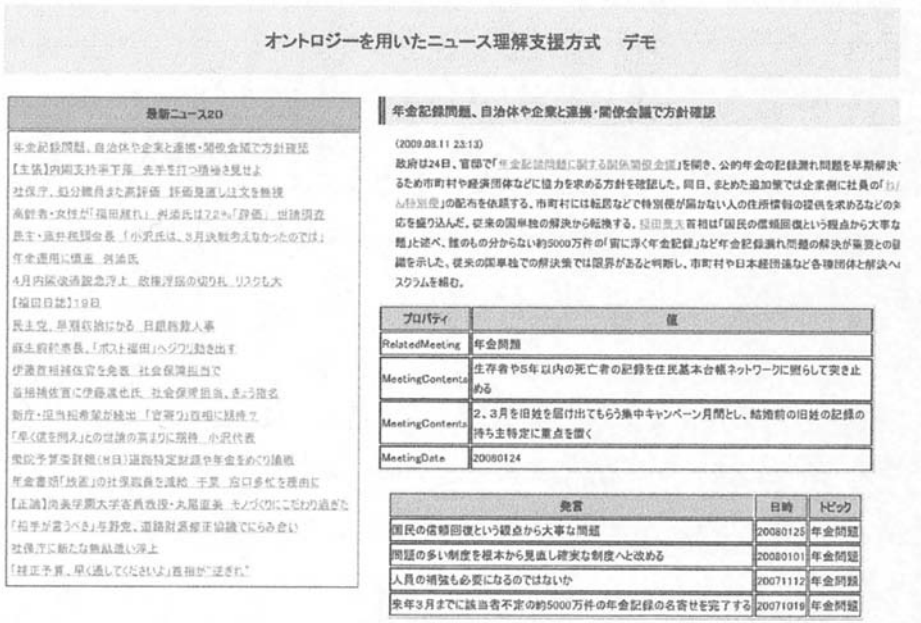

図 6: 実装画面例

予めアノテーションがされていないと処理を行えないという問題点と，コンテンツ提供者がアノテーションを付与しなければならないという問題点が存在する．しかしオントロジーを用いる場合，文書が事前にアノテーションの付与がされていなくとも処理ができる点と，コンテンツ提供者はアノテーションを意識する必要がない点で有用性がある．

4.2.2　オントロジーの構築的評価

構築したオントロジーの評価を行う．本稿では人物の発言と記事トピックを ex:Related プロパティで紐付けすることで関係を示した．そのために記事トピックに関連する発言に限定して抽出することができる．

上記の例によると，抽出された関連情報から 2007 年 11 月に福田総理が 5000 万件の宙に浮く年金記録の名寄せを完了すると名言していることがわかるが，2008 年 2 月 24 日のニュースにて記録漏れ問題の解決が重要との認識を示すなど，去年の見通しに対して作業が予定通り進んでいないことがわかる．このようにニュースを読むユーザの理解を支援する情報の提供ができたと言える．

しかし実際にユーザに使ってもらったりするなど，より詳細の評価が必要である点で課題が残る．

5　まとめ

ニュースは速報性が重視され古くなるほど価値が失われていくため，価値が失われる前に体系化し知識化することが必要であると考えた．そこでニュースオントロジーを構築し情報を保持することで，コンピュータが知識を処理できる体系作りを行った．本稿では今回構築したオントロジーを用いてニュースを閲覧するユーザの理解を支援する方式を提案した．

閲覧記事のトピックに関する動向を知らないユーザの理解を支援することができ，従来記事を読む際に理解できないまま読み飛ばしていた用語の理解も深まる．さらに二次的に検索をして理解をする手間を軽減することができる．

今後の展望として，まずオントロジー構築の

半自動化が必要である．今回は手作業にて構築を行ったため，非常に手間のかかる作業であり非効率的であった．また今回は人物の役職や発言を中心に関連情報として抽出したが，トピックに関して賛成的な発言なのか反対的な意見なのか，人物の発言に対して他人物がどんな発言をしているのかなど発言の比較を行うことでより詳細な関連情報を提供したいと考えている．

参考文献

[1] 吉田慶章, 柿崎淑郎, 辻秀一. タグクラウドを用いた注目情報提示方式. 第 69 回情報処理学会全国大会, 2007. 6S-2.

[2] Mills Davis. Semantic Wave 2008 Report : Industry Roadmap to Web 3.0 & Multibillion Dollar Market Opportunities. 2008. EXECUTIVE SUMMARY.

[3] 北山大輔, 角谷和俊. ニュースアーカイブのためのコンテンツ構成順序を用いた比較ニュース検索. 電子情報通信学会 第 18 回データ工学ワークショップ, 2007. DEWS2007 A9-4.

[4] 森純一郎, 松尾豊, 石塚満. Web からの人物に関するキーワード抽出. 人工知能学会論文誌, Vol. 20, No. 5, pp. 337–345, 2005.

[5] 奥田奈央, 難波英嗣, 奥村学. 新聞記事と blog からの動向情報の抽出と可視化. 言語処理学会 第 13 回年次大会, 2007.

[6] 綾聡平, 松尾豊, 岡崎直観, 橋田浩一, 石塚満. 修辞構造のアノテーションに基づく要約生成. 人工知能学会論文誌, Vol. 20, pp. 149–158, 2005.

[7] Tim Finin, James Mayfield, Anupam Joshi, R. Scott Cost, and Clay Fink. Information Retrieval and the Semantic Web. In *Proceedings of the 38th Hawaii International Conference on System Sciences*, 2005.

[8] 児玉政幸, 大圃忠親, 新谷虎松. メタデータを活用した NewsML マネジメントシステムの試作. 人工知能学会 第 20 回全国大会, 2006. 3B3-3.

[9] 橋田浩一, 和泉憲明. オントロジーに基づく知識の構造化と活用. 情報処理学会誌「情報処理」, Vol. 48, No. 8, pp. 843–848, 2007.

[10] 林良彦. セマンティック Web と言語資源・言語技術. 情報処理学会誌「情報処理」, Vol. 48, No. 8, pp. 857–863, 2007.

[11] 岡崎直観, 石塚満. 日本語新聞記事からの略語抽出. 人工知能学会 第 20 回全国大会, 2006. 2G4-4.

[12] 來村徳信, 鷲尾尚哉, 小路悠介, 溝口理一郎. 技術知識管理のための機能に関するオントロジーとセマンティックアノテーション. 人工知能学会 第 19 回全国大会, 2005. 2D1-03.