2E1-3

オントロジーを用いたニュース理解支援方式

A New Support Method for Understanding News Using Ontology

吉田　慶章[*1]　　柿崎　淑郎[*2]　　辻　秀一[*3]
Yoshiaki YOSHIDA　Yoshio KAKIZAKI　Hidekazu TSUJI

[*1]東海大学大学院　工学研究科
Graduate School of Engineering, Tokai University

[*2]東京理科大学　工学部
Faculty of Engineering, Tokyo University of Science

[*3]東海大学　情報理工学部
School of Information Science and Technology, Tokai University

There is service which supports for user's understanding the articles because user has a few knowledge sometimes. However it is difficult for user to understand the articles by using support service. In this paper, we propose a new support method which provides additional information for understanding the articles using ontology. We construct ontology with extracted knowledge from the articles. We implement this method, and discuss effectiveness by comparing our method with related ones.

1. はじめに

近年，ウェブのトレンドの変化から Web2.0 時代と言われている．Web2.0 の特徴としてユーザが積極的に情報を発信していくことが挙げられるが，その一方，多種多様な情報が氾濫してしまうという問題がある．

そこで，ウェブコンテンツにコンピュータ可読の意味を付与し知的なウェブを実現しようとするのが，セマンティックウェブの目指すビジョンの一つであり，次に来る Web3.0 時代の中心トレンドである．文献 [Davis 08] 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．

本稿では，このセマンティックウェブの基盤技術であるオントロジーを用いて，ユーザのニュース理解支援を行う方式を提案する．ここでユーザの理解支援とは，コンテンツをより深く，正しく知るために必要な情報を提供することで，ユーザの理解を促進させることを意味する．オントロジーを用いることで，語や物事の意味，関係をコンピュータ可読の形式で明確化することが可能となり，理解支援を行うために有用であると考えた．本稿では理解支援を行う対象をニュースとした．

2. 従来方式

2.1 従来サービス

ニュースは常にトピックの最新情報を伝えるものであるため，閲覧記事に関する過去の報道や動向を把握していない場合，記事を理解できない可能性がある．

そこで，ウェブ上ではユーザのニュース理解を支援するサービスが行われている．その一例として MSN 産経ニュース[*1] があり，このサービスでは，記事中のキーワードから記事内検索・ウェブ検索・画像検索を行うことができる．記事内検索を利用し同トピックの過去の記事や類似の記事を読むことは，記事をより深く知ることができる点で理解の支援となっている．

連絡先: 吉田慶章, 東海大学工学研究科, 〒 259-1292　神奈川県平塚市北金目 1117 番地, 0463-58-1211(内線:4104), y.yoshida22@gmail.com

∗1　http://sankei.jp.msn.com/

しかし MSN 産経ニュースの問題点として，必ずしも理解支援となる情報が提供されてないことが挙げられる．実際にあった上記の例として，年金問題に関する記事中の福田康夫というキーワードから外国人参政権付与に関する記事が提供されることがあった．この例は，記事のトピックには着目せず福田康夫というキーワードにのみ着目し関連する記事を提供するために起こると予想できる．

そしてウェブニュースの特徴として，記事の公開から二週間から一ヶ月程度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．そのため記事内検索の対象は比較的新しいニュースのみに限定されてしまう点も問題である．

2.2 関連研究

北山らの研究 [北山 07] 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，最適なニュースコンテンツの提供を行っている．ニュース理解を促進させるコンテンツを提供するという手段は類似していると考えられるが，本稿ではコンテンツを記事ではなく，記事から抽出した知識としている点で異なっている．

綾らの研究 [綾 05] では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．このように文書にアノテーションを行い検索や要約，テキストマイニングに利用する試みが多く行われている．本稿ではアノテーションではなくオントロジーを用いる点で異なっており，アノテーションとオントロジーの大きな違いはその特徴である．アノテーションでは各コンテンツ提供者が各々の意思で付与するために，意味の共通化が図れない場合が多くあるが，オントロジーを用いることで意味付けの段階を一元化することができ，より有用である．またオントロジーは情報の意味や概念，関係を詳細に体系付けることができる点でも利点がある．

3. ニュース理解支援方式

本稿ではニュース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．

本稿において，知識とはオントロジーが保持する体系化された情報，理解支援情報とはニュース記事の理解を促進させる

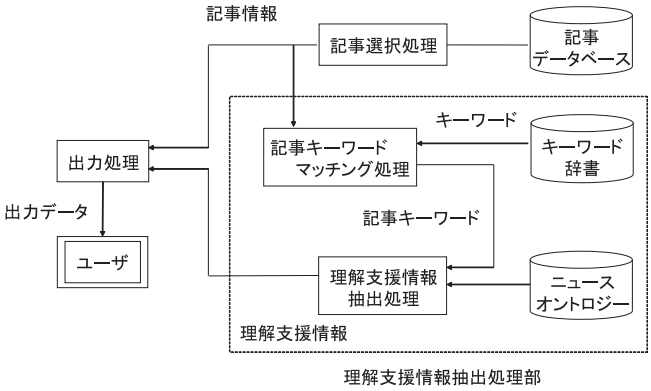

図 1: ニュース理解支援方式の概要図

情報と定義する．またユーザとは本方式を利用してニュース記事を閲覧する人と定義する．

図 1 に本稿で提案するニュース理解支援方式の概要図を示す．まず記事データベースから見出し・本文・日時・記事トピックを含んだ記事情報を抽出する．記事キーワードマッチング処理では MeCab[*2] を利用し記事情報とキーワード辞書に共通に含まれるキーワードを抽出し，記事キーワードとする．キーワード辞書とはオントロジーが保持するクラス名やインスタンス名をキーワードとした MeCab 用の辞書データ (.csv) である．理解支援情報抽出処理では記事キーワードと記事トピックに適した理解支援情報をオントロジーから抽出する．出力処理では受け取った記事情報に理解支援情報を付与しユーザに提供する．

MSN 産経ニュースの問題点として挙げた理解支援となる情報が提供されてないことに対して，本稿では各記事に関連する理解支援情報をオントロジーより抽出しユーザの提供することで解決する．

またコンテンツとして記事を提供していた MSN 産経ニュースや北山らの研究 [北山 07] とは異なり，理解支援情報を提供することで二次的に記事を読む手間がなくなる．

3.1 ニュースオントロジー

本稿で構築したニュースオントロジーは，ニュース記事中に表れる用語や重要な情報を体系化し，関連付けた知識を指す．今回体系化する知識のドメインは政治ジャンルに限定し，ニュース記事やウェブサイトから抽出した．またニュースオントロジーは OWL 言語を用い，protege[*3] により手動で構築した．

図 2 にニュースオントロジーの例として，現総理大臣である福田康夫を取り巻く知識を示す．楕円はクラス，四角はインスタンス，二重線で囲まれた四角はリテラルである．

福田康夫は People クラスのインスタンスとして生成している．そして内閣府から役職である内閣総理大臣をインスタンスとして生成している．人物が現在その役職に就任していることを表す `ex:CurrentAssumption` プロパティを用いて福田康夫と内閣総理大臣を連結することで，就任関係を示している．

`ex:Saying` プロパティは人物が発言したことを表す．発言クラスのインスタンスとして生成された二個の発言が福田康夫から `ex:Saying` プロパティで連結されていることから，これらの発言は福田康夫のものであることがわかる．

*2 http://mecab.sourceforge.net/
*3 http://protege.stanford.edu/

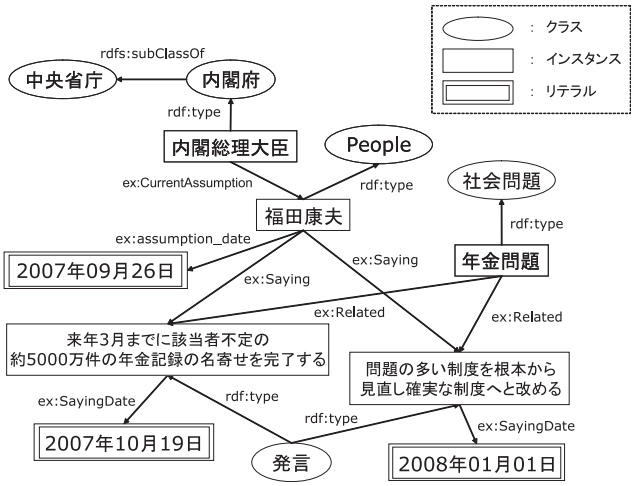

図 2: ニュースオントロジー例

```
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 3: SPARQL クエリ例

`ex:Related` プロパティはそれぞれの発言がどのニューストピックに関連するかを表す．社会問題クラスのインスタンスとしてニューストピックが生成され，その一例として年金問題がある．この年金問題と二個の発言が `ex:Related` プロパティで連結されていることから，これらの発言は年金問題に関連するものであることがわかる．

`ex:AssumptionDate` プロパティは人物が現在の役職に就任した日付を表し，`ex:SayingDate` プロパティは各発言がされた日付を表している．

3.2 オントロジーからの理解支援情報抽出

理解支援情報抽出には SPARQL (SPARQL Protocol and RDF Query Language)[*4] を用いた．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．

図 3 に例としてニュースオントロジーから福田康夫の年金問題の記事に関する理解支援情報（ここでは発言）を抽出する SPARQL クエリを示す．

3 行目では `ex:Saying` プロパティを用いて福田康夫の発言を`?say` にバインドしている．4 行目ではこのバインドされた発言の発言日を `ex:SayingDate` プロパティで`?s_date` に，5 行目では発言がどのニューストピックに関連しているのかを `ex:Related` プロパティで`?s_relate` にそれぞれバインドしている．6 行目では `FILTER` を用いることで，`?s_relate` にバイ

*4 http://www.w3.org/TR/rdf-sparql-query/

表 1: SPARQL クエリ結果例

?say	?s_date	?s_relate
国民の信頼回復という観点から大事な問題である	20080125	年金問題
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	20080101	年金問題
人員の補強も必要になるのではないか	20071112	年金問題
来年 3 月までに該当者不定の約 5000 万件の年金記録の名寄せを完了する	20071019	年金問題

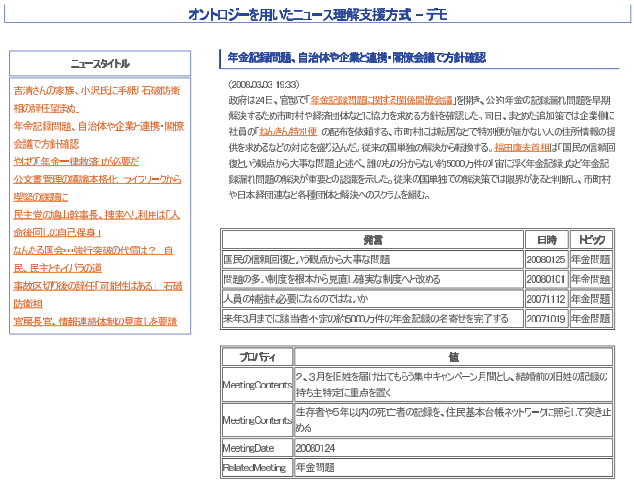

図 4: プロトタイプシステムの画面例

ンドされたニューストピックが年金問題であるもののみを抽出することができる．また，8 行目の ORDER BY では抽出された発言を発言日でソートを行っている．時系列にソートすることで，発言の変化が明らかになり記事理解が促進されると考える．

最終的にこのクエリ例から抽出できる理解支援情報は，数ある福田康夫の発言の中から，年金問題に対する発言のみに限定されたものである．よって年金問題の記事に対して不要な理解支援情報は提供されないことがわかる．

そして表 1 が図 3 のクエリの結果例である．三つの発言は全て福田康夫のものであり，それぞれ年金問題に関係している．

他のクエリを投げることで人物の役職歴や掲げているスローガンなどを抽出することができる．また人物だけではなく，例えば「年金問題」のようなキーワードからはそのニュースの動向を理解支援情報として抽出することができる．

4. 試作と評価

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システムの試作を行った．プロトタイプシステムの画面例を図 4 に示す．

左側にはニュースタイトルが並び，各タイトルを選択することで対応した記事データが右側に出力される．この出力データは，図 1 で示した記事情報に理解支援情報を付与したものである．PHP によりオントロジー処理には RAP - RDF API for PHP V0.9.5*5 を利用している．

*5 http://www4.wiwiss.fu-berlin.de/bizer/rdfapi/

4.2 アンケート評価

提案方式の有効性を検証するため，今回研究室の学生 10 名にアンケートを依頼した．設問項目を以下に記す．

設問 1 理解が深まりましたか？(有効性評価)

設問 2 提供された知識は十分でしたか？(網羅性評価)

設問 3 他にどんな情報が提供されると理解が深まると思いますか？

設問 4 不要と感じる情報はありましたか？

設問 5 その他

今回は被験者に対してイージス艦衝突事故に関するニュースを提供した．ユーザにはシステムを使う前にイージス艦衝突事故に対する理解レベルを以下の 4 項目から選択してもらった．A を選択したユーザは有識者と見なし，網羅性評価を含む設問 2,3,4,5 を回答してもらい，B,C,D を選択されたユーザには，有効性評価を含む設問 1,3,4,5 を回答してもらった．

A 詳しく知っている

B 多少知っている

C あまり知らない

D このニュースは聞いたことがない

表 2 に設問 1 と設問 2 の結果を示した．回答人数 10 人のうち，理解レベル A を選択したユーザが 3 人，理解レベル B,C,D を選択したユーザが 7 人だった．設問 1 に関して 6 人が「少し深まった」と回答し，1 人が「あまり深まらなかった」と回答した．設問 2 に関してそれぞれ 1 人ずつが「十分だった」，「ある程度十分だった」,「不足している」と回答した．

表 3 に設問 3，設問 4，設問 5 の結果の一部を示した．設問 3 では，発言の比較を行う方が有効ではないかという意見をもらった．本稿では人物の発言を単に提供するだけであるので今後の課題とする．また，設問 4 の「理解支援になるとは思えないような発言がある」，設問 5 の「情報が不十分な場合，理解を妨げる可能性がある」など，提供する情報の網羅性や質には課題が残ることがわかった．

これらの結果から提供する理解支援情報に関して再検討が必要であるが，提案方式の有効性は確認することができたと言える．

5. まとめ

ニュースを理解するにはニューストピックの過去の報道や動向を把握している必要がある．そこで本稿では構築したニュースオントロジーを用いて，ユーザの記事に適する理解支援情

表 2: 設問 1，設問 2 のアンケート結果

設問 1		設問 2	
回答	回答数	回答	回答数
とても深まった	0	十分だった	1
少し深まった	6	ある程度十分だった	1
あまり深まらなかった	1	不足している	1
何も深まらなかった	0	大変不足している	0

表 3: 設問 3，設問 4，設問 5 のアンケート結果

設問	コメント
設問 3	発言の比較を行うとより理解が深まるのではないか 画像が表示されるとイメージしやすくなるのではないか
設問 4	提供される知識が長すぎる 理解支援になるとは思えないような発言がある
設問 5	情報が不十分な場合，理解を妨げる可能性がある

報を提供することでニュース理解を支援する方式を提案した．そしてアンケート評価の結果から，本方式で理解を深めることができたとの声を得ることができ，本方式の有効性を確認できた．

今後の展望として，まずオントロジー構築の半自動化が挙げられる．今回はオントロジー構築を手動で行ったために，非常に手間のかかる作業であり非効率的であった．オントロジーの自動構築における先行研究 [内田 03, 中山 08] などを参考に，ニュースを対象としたオントロジー自動構築の方式を確立する．

また，アンケートの結果から得られた提供する理解支援情報の網羅性や質に関する課題に対して，オントロジーを自動化することで網羅性の向上には期待できると予測できるが，質の向上に関しては，各理解支援情報を重要度に応じてスコアリングするなどして解決していきたいと考える．

参考文献

[Davis 08] Davis, M.: Semantic Wave 2008 Report : Industry Roadmap to Web 3.0 & Multibillion Dollar Market Opportunities (2008), EXECUTIVE SUMMARY

[北山 07] 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

[綾 05] 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)

[奥田 07] 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)

[森 05] 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)

[数原 08] 数原 良彦, 戸田 浩之, 櫻井 彰人：ブログ記事を用いた複数話題語間の動作関係抽出手法, 電子情報通信学会論文誌, Vol. J91, No. 3, pp. 619–627 (2008)

[中山 08] 中山 浩太郎, 原 隆浩, 西尾 章治郎：自然言語処理とリンク構造解析を利用した Wikipedia からの Web オントロジ自動構築に関する一手法, 電子情報通信学会 第 19 回データ工学ワークショップ (2008), DEWS2008 A3-2

[内田 03] 内田 英里, 石野 武志：オントロジーの自動構築に関する基礎的研究, 人工知能学会第 3 回セマンティックウェブとオントロジー研究会 (2003), SIG-SW&ONT-A301-05

[白井 07] 白井 清昭, 徳永 健伸：呼応する名詞の包含関係に着目した助数詞オントロジーの自動構築と評価, 情報処理学会 第 181 回自然言語処理研究会, pp. 127–134 (2007), NL-181(20)