語句間の意味的リレーションに基づくキュレーションエージェント

横尾　亮平[†a)]　川村　隆浩[†,††]　清　雄一[†]　田原　康之[†]
大須賀昭彦[†]

Curation Agent Based on Semantic Relations between Events

Ryohei YOKOO[†a)], Takahiro KAWAMURA[†,††], Yuichi SEI[†], Yasuyuki TAHARA[†], and Akihiko OHSUGA[†]

あらまし　近年，ユーザの行動やソーシャルメディア上での発言を興味・関心として分析し，ニュース記事を推薦するキュレーションサービスが普及している．膨大な情報から自分で必要なものを探さなくても，自身の興味に沿った情報が手に入ることで利用者が増加している．既存のコンテンツベースの情報推薦システムに関する研究では記事推薦のために各語句を特徴としているが，頻出する語句を重要視しており語句間の関係を特徴として用いていない．本研究は，ユーザが興味・関心を示す記事に表れる語句間の意味的リレーションを用いることで，ユーザにとって興味度の高いニュース記事を収集・推薦するキュレーションエージェントを提案する．語句間の意味的リレーションは Linked Data で表現する．ユーザが興味・関心を示す記事文章からインターネット上のニュース記事を推薦する手法を提案する．評価実験の結果，興味度の指標の平均値は 4 点満点中 3.07 であり，本手法はユーザに対して興味度の高いニュース記事を推薦できるキュレーションエージェントとして有効であることが明らかになった．

キーワード　Web エージェント，Linked Data，Semantic Web，情報推薦，情報検索

1. まえがき

近年，Gunosy(注1) や Vingow(注2) などのニュース記事を自動で収集し，推薦するキュレーションサービスが普及し，膨大な情報から自分で必要なものを探さなくても，簡単に関心に合う情報が手に入るようになりつつある．多くは，ソーシャルメディアである Facebook やマイクロブログの Twitter と連携利用することで取得したユーザの発言から興味・関心を抽出，またサービス上でのユーザの記事閲覧履歴を学習することで，ユーザに最適な情報を配信している．キュレーションサービスの出現を背景に，ニュース記事は必要な情報をユーザが能動的に取得する存在から，ユーザへ自動で配信される存在へ形を変えつつある．

本研究では，ユーザが興味・関心を示すニュース記事文章内の語句間の意味的リレーションに着目し，ユーザにとって興味度の高いニュース記事を推薦することを目指す．本論文では語句間の意味的リレーションを Linked Data で表現する．インターネット上から取得したニュース記事群とユーザが興味・関心を示すニュース記事群からそれぞれ Linked Data を構築する．二つの Linked Data 間の類似する部分グラフを用いることで，ユーザに興味度の高いニュース記事を提供するキュレーションエージェントを提案する．例として，“クリミアの美人すぎる検事総長” [1] というニュースが存在する．昨今，インターネット上やマスコミの報道で美人すぎる市議や美人すぎる海女といった記事の出現により，美人すぎる○○という言葉が生まれ，興味・関心が集まっている．これには，「美人 → (職業) → 職業名」という意味的リレーションが存在する．美人と意外な組み合わせの検事総長という語句

† 電気通信大学大学院情報システム学研究科，調布市
Graduate School of Information Systems, The University of Electro-Communications, 1–5–1 Chofugaoka, Chofu-shi, 182–8585 Japan
†† (株) 東芝研究開発センター，川崎市
Corporate Research & Development Center, Toshiba Corp., Komukaitoshibacho 1, Kawasaki-shi, 212–0001 Japan
a) E-mail: r-yokoh@ohsuga.is.uec.ac.jp
DOI:10.14923/transinfj.2014SWP0017

(注1)：http://gunosy.com/
(注2)：https://vingow.com/

の意味的リレーションから，クリミア美人検事総長に興味の矛先が向けられた．それまでは日本であまり知られていない存在であったクリミアという地名にも興味を引く結果となった．このことは他国の美人検事やクリミアでは他にどのような美人が存在するかなど，意味的リレーション「美人 → (職業) → 検事総長，検事総長 → (地名) → クリミア」により連想される他ニュースへも多くの関心が集まると想像される．

文章表現に Bag-of-Words ベクトルを用いる既存手法の場合，ユーザが興味 · 関心を示す文「政情不安の続くクリミアにおいて美人すぎる検事総長が大人気」から三つのユーザの興味 · 関心語を抽出すると，ユーザが興味 · 関心を示す正しい語句の組み合わせが「美人，検事総長，クリミア」であっても,「政情不安，続く，大人気」となりえる．また，正しくユーザの興味 · 関心が抽出できた場合でも，1 文中に 3 語句が全て出現する他ニュース記事は非常に限られる．語句間の意味的リレーションに着目すると，記事中の「美人-検事総長」の出現する文と「検事総長-クリミア」の出現する文とを繋げれば,「美人-検事総長-クリミア」という 3 語句が出現するニュース記事を探し出すことが可能である．ここで，記事全体から「美人，検事総長，クリミア」という 3 語句だけで探索すると，語句間の関係を特徴に用いていないので関係のないニュース記事まで探してしまう．そのため，語句間の意味的リレーションに基づきユーザの興味 · 関心事を Linked Data(注3) を用いて表現することにより，ユーザの興味 · 関心の具体化を試みる．Linked Data とはデータを再利用しやすいような形で構造化し，公開 · 共有するための Web 技術である．Linked Data に変換することで計算機が扱いやすい整理された情報として利用できる他，類似する部分グラフの検索が容易になる．

本論文は以下のように構成される．**2.** にて関連研究を紹介する．**3.** で提案手法の概要について述べる．**4.** で提案手法の有用性を示すための評価実験について示す．**5.** では，本研究のまとめと今後の展望を述べる．

2. 関 連 研 究

既存のコンテンツベースの情報推薦システムに関する研究 [2]～[4] では文章内の各語句が特徴語として用いられている．特徴語を使った文章表現に Bag-of-Words ベクトルが一般的に用いられている．Bag-of-Words モデルを用いた推薦システムでは tf-idf 法やページランク法，トピックモデル法といったアルゴリズムを用いて，頻出語の特徴語に重み付けをし，重み付けが大きい語句を含む文章を推薦する．ユーザの興味 · 関心を示す文章内で頻出する語句を含む文章が推薦されているが，語句間の意味的リレーションを用いてニュース記事を推薦しているわけではない．

音声対話システムの研究分野では，記事の文章を述語項構造解析を用いて情報を抽出し，述語項構造が類似する情報推薦や検索を行う手法が提案されている [5]．Bag-of-Words モデルと比較しても，より的確な応答生成が確認されている．

コンテンツベースの推薦においては語句の意味を特徴に適用した手法が存在する．概念辞書 WordNet(注4) や語彙意味構造辞書 VerbNet(注5)，そのコーパスである SemLink(注6) などの言語資源が用いられている [6]．Capelle ら [7] はユーザの嗜好する記事に含まれる語句を特徴に利用し，WordNet と検索エンジンの Bing を用いて推薦対象となる文章との語句類似度を算出している．そして，類似する語句をもつニュース記事の推薦を行っている．

また，Khrouf ら [8] はイベント情報サイトのメタ情報 (場所/時間/タグ/ジャンルなど) を Linked Data 化し，データ構造の類似度を用いた手法と協調フィルタリング手法とのハイブリットによりイベント情報推薦システムを構築している．Linked Data だけではなく様々な情報と組み合わせた手法が存在している [9] ことも含め，近年，Linked Data を用いた推薦システムが多く提案されている．また，データ間の類似度を図る指標にも様々な提案がされている [10]．しかし，文章内に存在する語句と語句間の意味的リレーションを Linked Data に変換して情報推薦を行っている研究はない．本論では語句間の意味的リレーションを特徴に利用することで Bag-of-Words モデルを用いた既存手法よりも，ユーザの興味 · 関心を具体化することができ，ユーザにより興味度の高いニュース記事を推薦することを示す．

(注3)：http://linkeddata.org/

(注4)：http://wordnet.princeton.edu/

(注5)：http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

(注6)：http://verbs.colorado.edu/semlink/

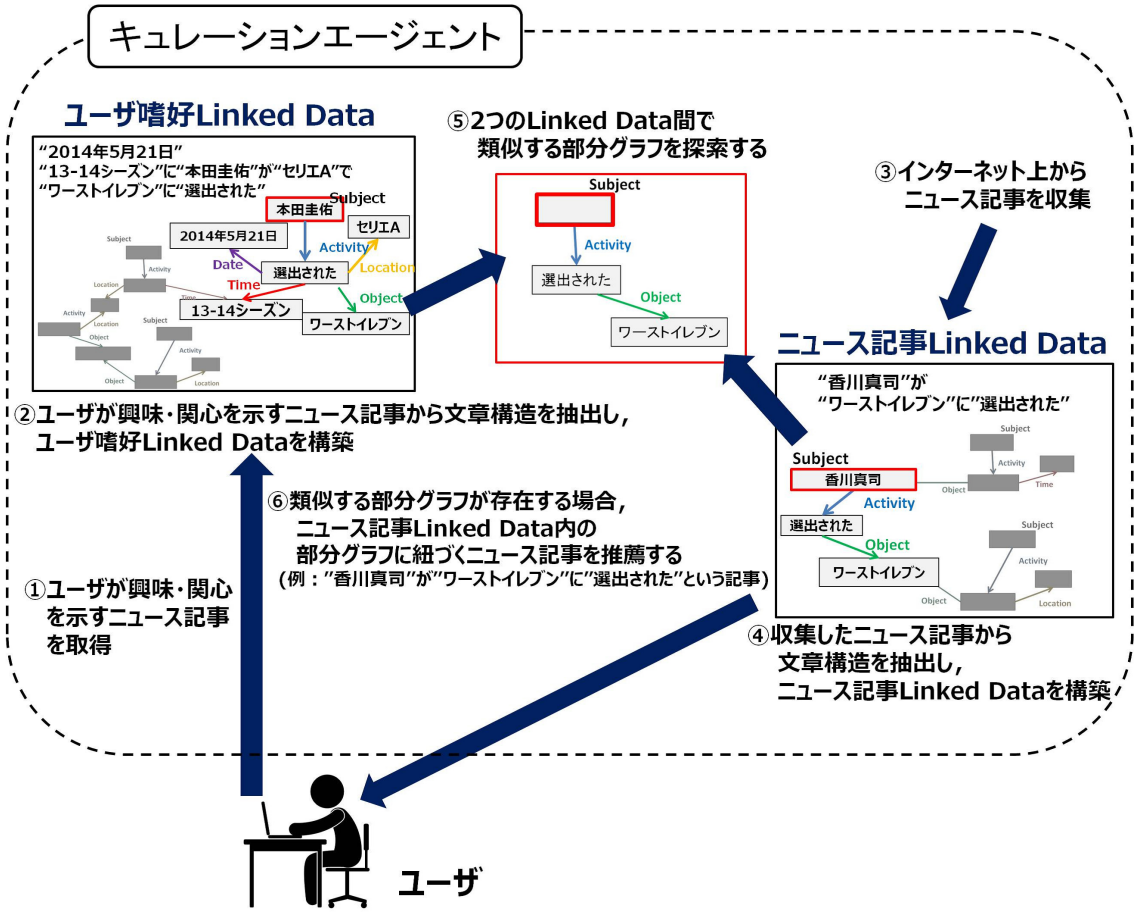

図 1　提案手法の概要
Fig. 1　Summary of our approach.

3. 語句間の意味的リレーションを用いた記事推薦手法の概要

本研究は，ユーザの興味・関心を示すニュース記事の文章構造を分析し，興味度の高いニュース記事をユーザに推薦することを目的とする．

提案手法の概要を図 1 に示す．(1)-まず，ユーザの興味・関心を示す一つのニュース記事を取得する．(2)-取得した記事を用いて，ニュース記事を構成する文章ごとにユーザ嗜好 Linked Data を構築する．Linked Data は記事の文章から抽出できる語句と意味的リレーションの組み合わせとする．(3)-続いて，ユーザに推薦するための記事をインターネット上から収集する．(4)-同様にニュース記事 Linked Data を構築する．(5)-二つの Linked Data 間で類似する部分グラフを探索する．(6)-もし，ユーザ嗜好 Linked Data と類似する部分グラフが News 記事 Linked Data に存在する場合はこの類似する部分グラフに紐づくニュース記事をユーザに推薦する．

本研究では二つの Property で繋がれた 3 語句ノードの部分グラフを利用し，類似する部分グラフを用いてのニュース記事検索をしている．文中のいずれかの箇所から 3 語句を拾うのではなく，類似する部分グラフの繋がりがある 3 語句を拾うほうが，より「関連度」の高い記事を選ぶことが出来，且つその内の一箇所 (1 語句) を変数として部分グラフマッチを行うことで，関連度を維持しながら,「興味度」の高い記事を推薦できると想定した．これはユーザの興味は関連度の高い記事の隣にある（元々，興味のある内容に近く（関連度が高く），わずかに異なる内容であることが定

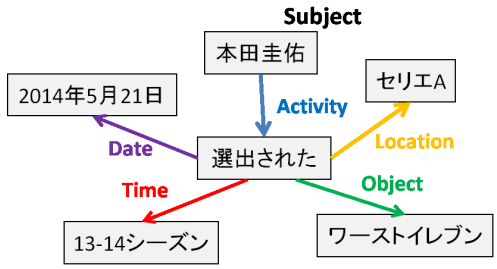

図 2　文章構造の例
Fig. 2　Example of sentence structure.

番である）と考えたからである．例えば，元々，クリミアの美人検事に興味がある人ならば，他の国の美人検事についても同様に興味があるだろうといえる．提案手法はこれを実現するための手法である．提案手法の詳細として，記事からの Linked Data 構築に関しては **3.1** に，類似する部分グラフ検索の手法と例を **3.2** にそれぞれ示す．

3.1　Linked Data の構築

3.1.1　文章構造の定義

Linked Data は一つの文から生成されたトリプル集合またはその部分集合を指す．Linked Data を構築するために，ニュース記事文章の文章構造を抽出する．記事文章内で出現する事象が表れる語句を対象とし，組み合わせたものを文章構造とする．Nguyen ら [11] は，Twitter や Web ページから得られる情報リソース内の文章で表現されている行動意図を認識するため，行動属性が表す語句を取得している．行動属性は (Who:だれが，When:いつ，Where:どこで，What:何を，Action:行動) のように定義されている．本研究では対象をマスメディアに掲載されているニュース記事に限定するため，新しく事象の定義を行った．Nguyen らの行動属性定義を参考に，Linked Data を構成する属性を以下に定義する．

- Subject (主語)
- Activity (主語の動作)
- Object (動作の対象)
- Date (日付)
- Time (時間)
- Location (場所)

例として，「2014 年 5 月 21 日 13-14 シーズンに本田圭佑がセリエ A のワーストイレブンに選出された」という文章を本文にもつニュース記事を取得した場合，抽出した文章構造は図 2 のように表現される．また，Linked Data 上では文章構造は複数のトリプルがリンクして構成される．トリプルとは「選出された (Activity) → Property:Object → ワーストイレブン (Object)」のように「主語 (Subject) → 述語 (Property) → 目的語 (Value)」の三つの要素でリソースに関する関係情報を表現しているメタデータモデルである．

3.1.2　前　処　理

ニュース記事の文内には，読者の目を引くためにカギ括弧 (「」, 【】, [ ] 等) や丸括弧が頻出する．文章構造が複雑になるため，抽出が難しくなる．抽出精度向上のために事前処理を適用した．カギ括弧に対する事前処理では文章構造を単純化するために「括弧内の文字列」と「括弧外の文字列」を分割する．また，丸括弧は略語の注釈，引用元の情報，著者名など，ほとんどの情報が **3.1.1** で定義した Linked Data を構成する属性として扱う必要がないため括弧内の文字列ごと除去した．

3.1.3　CRF による文章構造の抽出

ニュース記事本文から文章構造を自動抽出するために，John D. Lafferty ら [13] が提案した CRF (Conditional Random Field) を利用した．CRF は系列ラベリング問題を解くことができ，重複する特徴をモデルに組み込むことができる識別モデルである．通常の識別モデルとは異なり，出力が出力集合の部分集合ではなく，系列となる特徴がある．形態素解析，品詞タグ付与などの系列ラベリング問題に利用されている．同じく CRF を用いて，ニュース記事や twitter などの文章から動作を表す語句として事象の抽出を行った越川ら [12] の研究に基づき，学習モデルを構築し，ニュース記事から文章構造を抽出する．

CRF を用いたニュース記事からの文章構造の抽出のために，本研究で対象とするデータセットの全文章に素性を付与し，CRF による学習のためにフォーマット変換を適用し，訓練用データ，テスト用データを作成する．素性には文章の文脈情報を表現した文脈 ID，係り受け情報の示す係り受け先の文脈 ID，品詞細分類系を表す品詞 ID を用いる．文脈 ID と係り受け先の文脈 ID は日本語係り受け解析器 Cabocha(注7)，品詞 ID は日本語形態素解析エンジン Mecab(注8) からそれぞれ取得する．

(注7)：https://code.google.com/p/cabocha/
(注8)：http://mecab.googlecode.com/

表 1　訓練データの概要
Table 1　Dataset for training.

メディア名	文の数	語句数	ラベル数	Subject	Activity	Object	Date	Time	Location
朝日新聞デジタル	98	2554	2296	473	705	891	93	58	76

表 2　文章構造ラベルの推測精度
Table 2　Accuracy of labeling.

	Subject	Activity	Object	Date	Time	Location	Weighted Average
Precision	76.92%	90.80%	87.52%	68.42%	48.03%	61.31%	85.07%
Recall	79.86%	88.36%	76.37%	87.05%	81.48%	83.33%	83.33%
F-measure	78.36%	90.26%	81.56%	76.62%	60.44%	70.64%	84.19%

文脈ID (Context ID)	係り受け先ID (Dependency ID)	表層系 (Surface)	品詞ID (POS ID)	属性ラベル (Event Label)
0	1	2014	48	B-Date
0	1	年	53	I-Date
1	2	5	48	I-Date
1	2	月	38	I-Date
2	3	21	48	I-Date
2	3	日	53	I-Date
3	7	13	48	B-Time
3	7	-	36	I-Time
3	7	14	48	I-Time
3	7	シーズン	38	I-Time
3	7	に	13	O
4	7	本田	43	B-Subject
4	7	圭	44	I-Subject
4	7	佑	44	I-Subject
4	7	が	13	O
5	6	セリエ	38	B-Location
5	6	A	38	I-Location
5	6	で	24	O
6	7	ワースト	38	B-Object
6	7	イレブン	38	I-Object
6	7	に	13	O
7	-1	選出	36	B-Activity
7	-1	さ	31	I-Activity
7	-1	れ	32	I-Activity
7	-1	た	25	I-Activity

図 3　訓練データの例
Fig. 3　Example of training data.

また，CRF を用いた特徴モデル生成のために訓練用データを作成する．フォーマット変換を行った文章に対して，**3.1.1** で定義した属性をラベルとして人手で付与する．訓練データの例として，属性ラベルを付与した文章「2014 年 5 月 21 日 13-14 シーズンに本田圭佑がセリエ A のワーストイレブンに選出された」を図 3 に示す．B はチャンクの先頭，I は内部，O は外部をそれぞれ示す．チャンクは文章内の語句の表層系をまとめた固まりを示す．構築したモデルを利用して，テスト用データの属性ラベル抽出を行う．出力は CRF により連続する同一の属性ラベルが付与された表層系のチャンクとする．

本研究では，インターネット上のニュースメディアである朝日新聞デジタル(注9)の記事のうち 10 月 3 日に掲載された 13 件 98 文章の日本語ニュース記事を取得し，訓練用データに利用した．訓練データの概要を表 1 に示す．10 交差検定により算出した文章構造の各属性ラベルの推測精度の平均と全てのラベルに対する推測精度を示す加重平均を表 2 に示す．

10 月 3 日に掲載された朝日新聞デジタルの記事 13 件に含まれる 98 文章を対象としたところ，図 2 のように全て正しいラベリングができている文章は 10 件であった．しかし，提案手法では正しいトリプル集合を完全に取得できていなくても，部分的にトリプルが取得できていれば，類似文章構造検索を行い，類似部分グラフを取得することができる．本研究では二つの Property で繋がれた 3 ノードで構成される 2 トリプルで類似する部分グラフを探索し，類似する部分グラフを用いてニュース記事を検索している．そのため，2 トリプルをもつニュース記事が存在していれば類似部分グラフ検索は可能である．ただし，現在筆者らが定義する Linked Data のスキーマの場合は Property:Activity のトリプルが取得できない場合は類似文章構造検索を行い，類似部分グラフの取得ができない．

3.2　類似部分グラフの検索

ニュース記事を推薦するために，少なくとも一つ以上の完全トリプルをもつ部分グラフを類似する部分グラフと定義する．完全トリプルの主語または目的語を含む別トリプルが両部分グラフに存在し，尚且つ該当主語または目的語がもつ述語も一致するトリプルが存在する部分グラフを探索する．ユーザ嗜好 Linked Data の部分グラフと類似するニュース記事 Linked Data 内の部分グラフに紐づくニュース記事をユーザに推薦する．提案手法の概要図 1 においてユーザ嗜好 Linked Data 内の “本田圭佑 (Subject)” を主語とする部分グラフとニュース記事 Linked Data 内の “香川真司 (Subject)” を主語とする部分グラフの類似を例とする．二つの部分グラフはトリプル「選出された (Activity) → Property:Object → ワーストイレブン (Object)」をもつため，完全なトリプルであるといえる．また，完全トリプルの主語である「選出された (Activity)」を

(注9)：http://www.asahi.com/

Algorithm 1 類似部分グラフ検索 Search Subgraph

```
Input: UserGraph, NewsGraph
Output: All_Subgraph
function PartialSearch(u_graph, n_graph)
    for all u_triple ∈ u_graph do
        for all n_triple ∈ n_graph do
            if PartialMatch(u_triple, n_triple) then
                Push n_triple into array X
            end if
        end for
    end for
    return X
end function

function CollectSubgraph(news_triple, X)
    for all x ∈ X do
        Push news_triple + x
                    into array Subgraph
    end for
    return Subgraph
end function

for all user_triple ∈ UserGraph do
    for all news_triple ∈ NewsGraph do
        if SimTriple(user_triple, news_triple) then
            u_graph ← CollectGraph(user_triple)
            n_graph ← CollectGraph(news_triple)
            X ← PartialSearch(u_graph,n_graph)
            Push CollectSubgraph(news_triple, X)
                        into array All_Subgraph
        end if
    end for
end for
return All_Subgraph
```

Algorithm 2 一致トリプル検索 Search Triple

```
Input: u_triple, n_triple
Output: Bool
function SimWords(u_word, n_word)
    if u_word == n_word then
        return True
    end if
    if WordNet(u_word, n_word) then
        return True
    end if
    if Jaccard(u_word, n_word) ≥ 0.5 then
        return True
    end if
    return False
end function

function SimTriple(u_triple, n_triple)
    if u_triple.property != n_triple.property then
        return False
    end if
    if SimWords(u_triple.subject, n_triple.subject) then
        if SimWords(u_triple.value, n_triple.value) then
            return True
        end if
    end if
    return False
end function
```

含むトリプル「主語 → Property:Activity → 選出された (Activity)」で部分一致する．このとき，ニュース記事 Linked Data の該当トリプルの主語部分は「香川真司 (Subject)」である．以上から類似する部分グラフ香川真司 (Subject) → Property:Activity → 選出された (Activity) → Property:Object →ワーストイレブン (Object)」に紐づくニュース記事（この場合，「香川真司がワーストイレブンに選出された」）をユーザに推薦する．

Linked Data の類似部分グラフを取得するためのアルゴリズムを Algorithm 1 に示す．ユーザの嗜好 Linked Data 内のトリプル *user_triple* の集合を *UserGraph*，ニュース記事 Linked Data 内のトリプル *news_triple* の集合を *NewsGraph* とし，それぞれを入力値とする．まず，*user_triple* と *news_triple* が二つの Linked Data 間で一致するトリプルであるかどうかを *SIMTRIPLE* によりチェックする．*SIMTRIPLE* は Algorithm 2 で述べる．一致するトリプルであった場合，該当トリプルの主語または目的語を含むトリプル集合を *CollectGraph* により取得し，それぞれ *u_graph*，*n_graph* とする．*PartialMatch* により *u_graph*，*n_graph* 間で同一の Property を含むトリプルが存在する場合，*n_triple* を変数 x として，類似部分グラフ構築のために収集する．最後に，*n_triple* と変数 x を連結させて類似部分グラフを構築し，*Subgraph* を出力する．そして，出力した *Subgraph* に紐づくニュース記事をニュース記事 Linked Data から取得し，ユーザに推薦する．

3.3 Entity Linking

ユーザ嗜好 Linked Data の部分グラフと類似する部分グラフをニュース記事 Linked Data から検索するためには，通常，主語 (Subject)· 述語 (Property)· 目的語 (Value) の 3 要素が全て一致するトリプルを含む部分グラフを探索する必要がある．しかし，トリプルの主語 · 目的語の語句表層形が完全一致することに限定すると，一致するトリプル数は非常に少ないと予想される．また，探索機会損失に繋がり，ユーザに推薦されるべき記事が推薦されない問題にもつながる．

そのため，Linked Data の各ノード (Subject, Value) の語句に対して Entity Linking を行い，探索機会を増加させることを試みる．Entity Linking と

は文章中に現れる Entity (語句) への参照表現を認識し，参照表現辞書の該当する意味にリンクするタスクのことである．例えば「ワーストイレブンに選出された」という文章があるときに「選出された」は「選ばれた」,「選抜される」などと同様の参照表現をもつ．語句の表層形一致よりも多くの探索機会を得ることができる．Bunnescu [14] らの研究は Entity Linking において草分け的存在である．Bunnescu らは Wikipedia の記事間のハイパーリンク構造を用いることで，固有名を同定し，曖昧性を解消する手法を提案している．また，Hoffart [15] らは Linked Data の部分的なグラフ構造を利用する手法を提案している．

本研究での Entity Linking の手法には日本語 WordNet(注10) と最も基本的な手法である Jaccard 係数を利用して，文字列の類似度を算出する．Jaccard 係数は式 (1) により定義される．Jaccard 係数とは二つの文字の集合 A，B の共通要素の割合を表す．入力を二つの文字の集合とすると，出力の値域は 0 から 1 の間を示し，1 に近づくほど二つの文字列集合間の類似度が大きくなる．

$$Jaccard(A,B) = \frac{A \cap B}{A \cup B} \tag{1}$$

本論で示す「選出された」「選ばれた」「選抜される」は一文字ごとの共起関係しかないため，Jaccard 係数による Entity Linking は，特に「選出された」「選ばれた」「選抜される」等の活用や語尾変化が存在する動詞においては脆弱な手法である．そのため対象となる語句を，いったん，MeCab により形態素を基本形に戻し，活用や語尾変換を統一し，更に日本語 WordNet を参照し類似語を検索することで脆弱性を軽減している．日本語 WordNet を利用することで新たに，7346 件の語句組を検索することができた．日本語 WordNet を用いて類義語だと判定された上記 7346 件の語句組のうち，ランダムに 100 件の語句組を抽出して筆者らが確認したところ，意味が同様であると判断できた語句組は 80 件であった．日本語 WordNet で検索しきれなかった部分については Jaccard 係数を用いている．

Entity Linking による一致トリプルを取得するためのアルゴリズムを Algorithm 2 に示す．入力値をユーザ嗜好 Linked Data を構成する一つのトリプル *u_triple* とニュース記事 Linked Data を構成する一つのトリプル *n_triple* とする．このとき，トリプル *u_triple* とトリプル *n_triple* の Property は同一のものを扱う．それぞれのトリプルがもつ Subject 同士 (*u_triple.subject,n_triple.subject*), Value 同士 (*u_triple.value,n_triple.value*) の文字列の類似を完全一致，日本語 WordNet を参照することによる類似語検索による一致 *WordNet*，Jaccard 係数による一致 *Jaccard* の順に判断する．また，Jaccard 係数のしきい値を 0.5 とし，Subject または Value の片方でも下回るトリプルは一致していないとして類似部分グラフ検索に利用しない．

4. 評価実験

ユーザの興味度（推薦されたニュース記事は興味のもてる記事か）が高いニュース記事を推薦できることを確認することを目的とする．本手法によりニュース記事を検索し，被験者に推薦したうえで本手法の有効性を確かめる．また，指標としては興味度の値さえ高ければ良いといえるが，参考情報として関連度（推薦されたニュース記事が興味・関心を示したニュース記事と内容が関連しているか）も評価指標に加えた．

4.1 データセット

ニュース記事 Linked Data 構築のためにインターネット上のニュースメディアの朝日新聞デジタル，NHK NewsWEB(注11) からそれぞれ日本語文章のニュース記事を収集した．記事収集期間は 2014 年 10 月 3 日から 2014 年 12 月 5 日まで収集した全 26,925 件の記事をデータセットに利用する．ユーザ嗜好 Linked Data の構築には朝日新聞デジタルから収集したニュース記事 590 件を利用した．収集期間は 2014 年 12 月 6 日〜2014 年 12 月 8 日である．データセットの概要と Linked Data を構成するユニークなノード数を属性ラベルごとに表 3 に示す．朝日新聞デジタルの記事数に対するノード数が多いのは，NHK NewsWEB で扱っている記事より文章が長いからである．

本研究では照応・共参照には対応していない．文脈依存性の高い日本語記事では照応・共参照解析が必要となる場合ある．CRF モデル作成のために訓練データ作成の際に利用した 10 月 3 日の朝日新聞デジタルのニュース記事 13 件に含まれる 98 文章を対象とし，照応・共参照解析が必要となる箇所を探した．特に，ゼロ代名詞や人称代名詞，上位語，略称などによる照応を除く，指示詞による照応・共参照に絞って調査し

(注10)：http://nlpwww.nict.go.jp/wn-ja/

(注11)：http://www3.nhk.or.jp/news/

表 3　ニュース記事 Linked Data のデータセット
Table 3　Dataset for news article Linked Data.

メディア名	記事数	ノード数	ラベル計	Subject	Activity	Object	Date	Time	Location
朝日新聞デジタル	6801	23362	24202	7396	5015	9288	1138	815	550
NHK NewsWEB	8201	3563	3655	830	657	1685	174	156	153

表 4　ユーザ嗜好 Linked Data のデータセット
Table 4　Dataset for users' preference Linked Data.

メディア名	記事数	ノード数	ラベル計	Subject	Activity	Object	Date	Time	Location
朝日新聞デジタル	590	1619	1639	470	413	591	63	63	39

ユーザ嗜好Linked Dataの部分グラフ	ニュース記事Linked Dataの部分グラフ
同じウィルス→Activity→検出されました→Location→鹿児島県出水市	今シーズン最初のツルの飛来→Activity→確認されました→Location→鹿児島県出水市の出水平野
Jリーグ→Activity→送っている→Object→東南アジアに熱視線を	同じ病気→Activity→送る→Object→熱い視線
内閣府→Activity→発表した→Date→4日	安倍晋三首相の夫人の昭恵さん→Activity→披露した→Date→4日

図 4　類似部分グラフの例
Fig. 4　Example of common subgraph.

た．その結果，対象記事に含まれる 2554 語中，“それ” “その” “これ” “このうち” などの指示語の出現数は 8 回であった．“その”·“領収書” という連続する指示語句が記事文章に含まれている場合でも，本手法では CRF によるラベリングを用いることによ “その領収書” という形で語句を抽出することができる．そして，他の記事文章に含まれる “領収書” とマッチすることができる．指示語以外の，ゼロ照応等の照応 · 共参照現象についても今後調査する必要はあるものの，少なくとも指示語による照応 · 共参照に起因する問題については，上記のように緩和できると考える．ただし，提案手法の適用範囲を拡大するには照応 · 共参照解析の組み込みが必須であり，この点は今後の課題である．

4.2　実 験 概 要

予備実験を行ったところ，ニュース記事 Linked Data と類似する部分グラフをもつユーザ Linked Data に紐づくニュース記事は 590 件中 43 件であった．今回はこの 43 件の記事を各被験者に掲示する．類似部分グラフを検索する際に用いた一致するトリプル間の Jaccard 係数が一番高い記事 1 件を被験者に推薦する．ユーザに記事を推薦するために検索した二つ Linked Data 間の類似部分グラフの例 3 件を図 4 に示す．左列がユーザ嗜好 Linked Data の部分グラフを示し，右列のニュース記事 Linked Data 部分グラフは類似する部分グラフを示す．

4.3　実 験 手 順

評価実験ではまず，各被験者には掲示する 43 件の記事から被験者に興味 · 関心を示す 5 件の記事を選択させる．5 件の記事から類似部分グラフの検索により取得することができる記事 5 件を推薦する．次に被験者が推薦された各記事の内容に対して興味度，関連度の評価を行う．

関連度，興味度の各指標につき，そう思う (4 点)· ややそう思う (3 点)· ややそう思わない (2 点)· 思わない (1 点) の 4 点満点の評価で解答させた．被験者とした 8 人は全員電気通信大学の学生である．また，ベースライン手法として文章内の重要度の組み合わせからニュース記事を推薦する手法を用いて比較実験を行った．特徴としてユーザが興味を示す記事文章内の名詞から構成される Bag-of-Words モデルを用いている．本提案手法と同様に各被験者が興味 · 関心を示した 5 件の各記事文章からそれぞれ名詞を抽出した．単語の重要度を出現頻度から計算する tf-idf を用いた．提案手法と同じ 8 人の被験者を対象とした．算出した tf-idf スコア上位 3 語を含むニュース記事をニュース記事 Linked Data と同様のデータセットから検索した．同様に各被験者に興味 · 関心を示す記事と対応する 5 件のニュース記事を推薦し，評価をした．

表 5　比較実験の結果
Table 5　Experimental result.

	興味度	関連度
提案手法	3.07	2.60
ベースライン手法	2.72	2.60

4.4　実 験 結 果

提案手法とベースライン手法の評価結果の平均値を表 5 に示す．関連度においては，提案手法とベースライン手法ともに同じスコアだが興味度においては提案手法がベースライン手法を上回った．ベースライン手法は記事に含まれる tf-idf スコア上位語句を利用して

いるため，関連性の高いニュース記事の推薦精度が高いと考える．提案手法では関連度を維持したまま推薦された記事に対するユーザの興味度の高いニュース記事を推薦することができた．

5. むすび

本論文では，文章内の語句と語句間の意味的リレーションを Linked Data により表現し，ユーザの興味・関心と類似する部分グラフの検索を行うことでインターネット上のニュース記事をユーザに推薦する手法を提案した．本手法は，Bag-of-Words モデルではなく，語句と語句間の意味的リレーションを特徴として用いている．そのため，ユーザの興味をより具体化し，興味度が高いニュース記事を推薦することができる．しかし，Activity のラベルが付けられた「あった」や「行った」などといった，トリプル間では一致していても，単体では推薦する記事の内容をイメージしにくいような語句を含む類似部分グラフが頻出した．それでもベースライン手法と関連度は同じスコアだったが，より関連度の高いニュース記事をユーザに推薦するためにはトリプルの組み合わせによっては「あった」や「行った」といった語句を含む類似部分グラフを用いて検索するニュース記事の推薦優先度を下げる必要があると考えられる．そのため，よりニュース記事の内容をイメージしやすいような語句を含む性質をトリプル集合がもたなければいけない．これについては tf-idf スコアの高い語句を含むトリプルにより検索する類似部分グラフによりニュース記事を推薦することで解決する．また，ニュース記事を検索する際に用いる部分グラフを構成するトリプルの組み合わせや語句がどのようであれば，よりユーザの興味を惹くニュース記事を推薦することができるのかの検証する．そして今後の課題として，より正確にニュース記事の文章構造を抽出するため，照応・共参照解析へ対応する．更に，Web アプリケーション作成を通して，ユーザの興味・関心を学習し，語句と語句間の意味的リレーションを用いた，より高精度なキュレーションエージェントの実現を目指す．

謝辞　本研究は JSPS 科研費 24300005，26330081，26870201 の助成を受けたものです．本研究を遂行するにあたり，研究の機会と議論・研鑽の場を提供して頂き，御指導頂いた国立情報学研究所/東京大学本位田真一教授をはじめ，活発な議論と貴重な御意見を頂いた研究グループの皆様に感謝致します．

文　献

[1] JCAST ニュースビジネス&メディアウォッチ,「萌え化」クリミアの美人検事総長日本で大騒ぎに英 BBC もびっくり，JCAST, http://www.j-cast.com/2014/03/22199873.html?p=all, 参照 Aug. 27, 2013.

[2] 早川　豪，岡部　誠，尾内理紀夫，“Twitter を利用したソーシャルニュース記事推薦システム，” 情処学研報，データベース・システム研究会報告，2011-DBS-153(16), pp.1–4, 2011.

[3] W.-J. Lee, K.-J. Oh, C.-G. Lim, and H.-J. Choi, “User profile extraction from Twitter for personalized news recommendation,” Proc. 16th Advanced Communication Technology, pp.779–783, 2014.

[4] W. IJntema, F. Goossen, F. Fransincar, and F. Hogenboom, “Ontology-based news recommendation,” Proc. 2010 EDBT/ICDT Workshops, pp.16:1–16:6, 2010.

[5] 吉野幸一郎，森　信介，河原達也，“述語項の類似度に基づく情報推薦を行う音声対話システム，” 情処学研報，SLP 音声言語情報処理，2011-SLP-87(11), pp.1–6, 2011.

[6] 松林優一郎，岡崎直観，辻井潤一也，“自動意味役割付与における意味役割の汎化，” 自然言語処理，vol.17, no.4, pp.59–89, 2010.

[7] M. Capelle, F. Hogenboom, and A. Hogenboom, “Semantic news recommendation using WordNet and bing similarities,” Proc. 28th Annual ACM Symposium on Applied Computing, pp.296–302, 2013.

[8] H. Khrouf and R. Troncy, “Hybrid event recommendation using linked data and user diversity,” Proc. 7th ACM Conference on Recommender Systems, pp.185–192, 2013.

[9] V.C. Ostuni, T.D. Noia, E.D. Sciascio, and R. Mirizzi, “Top-N recommendations from implicit feedback leveraging linked open data,” Proc. 7th ACM Conference on Recommender Systems, pp.85–92, 2013.

[10] R. Meymandpour and J.G. Davis, “Recommendations Using Linked Data,” Proc. 5th Ph.D. Workshop on Information and Knowledge, pp.75–82, 2012.

[11] T.M. Nguyen, T. Kawamura, Y. Tahara, and A. Ohsuga, “Self-supervised capturing of users´ activities from weblogs,” Int. J. Intelligent Information and Database Systems, vol.6, no.1, pp.61–76, 2012.

[12] 越川兼地，川村隆浩，中川博之，田原康之，大須賀昭彦，“CRF を用いたメディア情報の抽出と LinkedData 化—ソーシャルメディアとマスメディアの比較事例，” 合同エージェントワークショップ&シンポジウム（JAWS2012）論文集，pp.1–9, 2012.

[13] J.D. Lafferty, A. McCallum, and F.C.N. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” ICML ’01 Proc. Eighteenth International Conference on Machine Learning, pp.282–289, 2001.

[14] R. Bunnescu and M. Pasca, “Using encyclopedic knowledge for named entity disambiguation,” Proc.

11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06), pp.9–16, 2006.

[15] J. Hoffart, M.A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum, “Robust disambiguation of named entities in text,” Proc. Conference on Empirical Methods in Natural Language Processing, pp.782–792, 2011.

（平成 26 年 8 月 27 日受付，12 月 17 日再受付，
27 年 3 月 5 日早期公開）

横尾　亮平

2013 年東京都市大学環境情報学部情報メディア学科卒業．2015 年電気通信大学大学院情報システム学研究科博士前期課程修了．

川村　隆浩

1992 年早稲田大学理工学部電気工学科卒業．1994 年同大大学院理工学研究科電気工学専攻修士課程了．同年，(株) 東芝入社．現在，同社研究開発センター主任研究員．工学博士．2001-2002 年米国カーネギ．メロン大学ロボット工学研究所客員研究員．2003 年より電気通信大学大学院情報システム学研究科客員准教授．2007 年より大阪大学大学院工学研究科非常勤講師．主としてマルチエージェントシステム，セマンティック Web の研究・開発に従事．情報処理学会会員．

清　　雄一　（正員）

1981 年生．2009 年東京大学大学院情報理工学系研究科博士後期課程修了．同年（株）三菱総合研究所入社．同社情報技術研究センター，金融ソリューション本部等に所属．2013 年より電気通信大学助教，現在に至る．分散コンピューティング，セキュリティ，プライバシー保護技術等の研究に従事．IEEE CS 会員．

田原　康之

1966 年生．1991 年東京大学大学院理学系研究科数学専攻修士課程修了．同年 (株) 東芝入社．1993～1996 年情報処理振興事業協会に出向．1996～1997 年英国 City 大学客員研究員．1997～1998 年英国 Imperial College 客員研究員．2003 年国立情報学研究所入所．2008 年より電気通信大学准教授．博士 (情報科学)(早稲田大学)．エージェント技術，及びソフトウェア工学などの研究に従事．情報処理学会，日本ソフトウェア科学会会員．

大須賀昭彦　（正員）

1958 年生．1981 年上智大学理工学部数学科卒．同年 (株) 東芝入社．同社研究開発センター，ソフトウェア技術センター等に所属．1985～1989 年 (財) 新世代コンピュータ技術開発機構 (ICOT) 出向．2007 年より，電気通信大学大学院情報システム学研究科教授．2012 年より，国立情報学研究所客員教授兼任．工学博士 (早稲田大学)．主としてソフトウェアのためのフォーマルメソッド，エージェント技術の研究に従事．1986 年度情報処理学会論文賞受賞．IEEE Computer Society Japan Chapter Chair，人工知能学会理事，日本ソフトウェア科学会理事を歴任．情報処理学会，電子情報通信学会，人工知能学会，日本ソフトウェア科学会，IEEE Computer Society 各会員．