特徴語と RDF を用いた情報推薦手法の提案

藤原哲[†1]　大場みち子[†2]　山口琢[†1]
奥野拓[†2]　伊藤恵[†2]

Web には数多くの情報が存在し，ユーザはその中で検索を繰り返すことで目的の情報を取得している．しかし，ユーザが目的とする情報に辿りつくためには高度な検索技術の活用や検索能力を求められるため，すべてのユーザが目的とする情報を取得することは容易ではない．そこで本稿では，ユーザが目的とする情報を推薦するための Web ページ情報の RDF 化手法の提案を行う．その中で，Web ページの文章から抽出する重要キーワードの RDF 化と閲覧中の Web ページ文章から抽出する重要キーワードを用いた情報検索・推薦手法の提案を行い，その有効性を確認するために行った実証実験について報告する．

Information Recommendation Method Using Feature Terms and RDF

TETSU FUJIWARA[†1]　MICHIKO OBA[†2]　TAKU YAMAGUCHI[†1]
TAKU OKUNO[†2]　KEI ITOU[†2]

A lot of information exists in the Web, and users get target information by searching from it. However, it is not easy for all of the users to get the target information because search capabilities and the use of advanced search technology is required. In this paper, propose a technique of description by RDF for Web page information to recommend information that users intended. We Propose an information search method and recommendation method using important keywords to be extracted from browsing Web page documents and description by RDF for important keywords extracted from the document of Web pages. After that, we report on the demonstration to confirm its validity.

1. はじめに

近年，インターネットの発達によりユーザは様々な情報を Web から取得することが可能となった．しかし，Web から情報を取得するためには，ユーザに高度な検索技術の活用や情報収集能力が求められるため，すべてのユーザが目的とする情報を取得することは難しい．現在，Web から情報を収集するための方法として，(1)キーワード検索，(2)リンクを組み合わせて情報収集を行うことが主流であるがそれぞれに次のような問題がある．

(1) キーワード検索

キーワード検索は，ページ内の文章に含まれるキーワードや Search Engine Optimization(SEO)によってメタデータ内に記述されたキーワードとユーザが入力したキーワードと一致させる最も一般的な検索方法であるが，次のような問題がある．まず，ユーザが情報収集対象とする物事の曖昧な知識や情報のみ持っている状態で検索を行う際に，的確なキーワードを知らなければ目的の情報まで辿りつきづらい．また，的確なキーワードを知っていてもヒットする Web ページが多く，多数にわたるページの閲覧や複数語検索(AND 検索)が必要となる場合，ユーザが目的を満たす情報を収集するまでに時間がかかる．さらに，多数のページを閲覧する際に有用な情報を見落すなどの問題がある．

(2) リンク

リンクとは，Web ページ内に他の Web ページの場所を指し示すことで，別な情報へユーザを遷移させるリンクという方法がある．キーワード検索でユーザが有用な情報を含むページを閲覧している場合に，更に詳細な情報や関連の深い情報を探索するのに有効である．しかし，ページ内で紹介されているトピックを構成する様々な要素すべてに対して，それぞれリンクが貼られていることは稀であり，情報発信者の一存で部分的なリンクの設定に留まっている．そのため，ユーザが求める情報に辿りつくリンクが Web ページ内に存在しない場合がある．また，リンクが存在していても古い情報へのリンクである場合や検索した時期に適した情報ではない可能性があるという問題がある．

以上の Web 検索における問題点を解決するアプローチとして，以下の２点の提案を行う．

1. 現在の検索では実現できない Web ページ内でキーワードが「含まれる」情報の検索ではなく，「重要である」情報の検索や時期等を踏まえた検索を行うため，Web ページ情報の構造化とデータ同士の関係性を構築する．

2. ユーザが着目している情報から更に詳細な情報や関連する情報を提示するために，閲覧中の Web ページ内のトピックを説明する文章から重要キーワードを抽出し，上記アプローチ 1 のデータに対する検索とリンクの推薦を行う．

本稿の構成は次の通りである．２章では関連研究と関連

†1 公立はこだて未来大学大学院
Graduate School of Future University Hakodate
†2 公立はこだて未来大学
Future University Hakodate

技術について述べる．３章では提案手法について述べる．４章では実装方式について述べる．５章では提案方式の評価実験について述べる．６章では本研究のまとめを述べる．

2. 関連研究/関連技術

本章では，2.1 節で Web ページを用いた情報推薦・検索支援に関する関連研究について述べる．2.2 節ではデータ間の関連を表し，関連の意味を表す Linked Open Data[1]を用いた情報推薦の関連研究について述べる．

2.1 Web ページを用いた情報推薦

ユーザの情報検索の支援を目的として Web ページの文章からキーワードを抽出する研究が，近年注目を集めている．例えば，渡辺らはスマートフォンブラウザでの検索支援を目的に，閲覧中の Web ページから文章を抽出し，文章から検索キーワード候補の抽出およびスコアリングを行い，提示することで検索の支援を行っている[2]．しかし，このアプローチではユーザに出力されるのは，そのページに存在する重要なキーワードである．そのため，そのキーワードを用いてユーザが検索を行なってもそのキーワードが重要である Web サイトが上位に来るとは限らないという問題がある.それは，現在のキーワード検索が検索されたキーワードを含むサイトを検索結果として扱い，ユーザが検索を行ったキーワードが文章内で重要ではなくても，閲覧数などによって上位に来るためである．また，ユーザに重要語が提示されたあとの検索が手動であるため手間がかかるという問題がある．

2.2 Linked Open Data (LOD)を用いた情報推薦

LOD を用いた新たな情報推薦の提案が近年多数報告されている．例えば，ユーザが閲覧中の Web ページに記述された文章と Resource Description Framework(RDF) [3]で表現されるリソース間の関連を用いて情報を推薦する研究がある[4]．ここでは，文章における最頻出語 1 つを「内容語」と呼び,「内容語」と関連が強い語を「関連語」とし，その 2 つの語の知識を探している．その知識を持つ別な情報を推薦することでユーザが閲覧中の Web ページに含まれる重要なトピックと関連がある情報を推薦している．しかし，この研究で用いられる最頻出語は,必ずしもその Web ページで扱っているトピックの重要なキーワードであるとは限らないため,文章によってキーワード抽出の精度が変わるという問題がある．また，文章中の最も重要なキーワードと位置付ける 1 つの情報のみを推薦しているため，検索支援が広範囲の情報を対象に行えないという問題がある．

LOD は, RDF を利用して関連するデータを利用するアークテクチャである．RDF で記述されるリソースの様々な関連に対して，意味を踏まえた検索クエリを与えることが可能となり，複雑でユーザの目的に合わせた検索が行える．次に，ユーザが関連する情報の取得範囲を限定することで，目的にあった情報を取得しやすくすることが可能になる．また，ユーザの目的に合わせて取得する情報の起点を定めることによって，ユーザの求める情報に辿りつく経路を選定しやすくすることもできるという利点がある．

本研究では, 閲覧中の Web ページにある文章から重要なキーワードを抽出し，そのキーワードを利用した検索も扱う．その際，現在の Web 検索ではデータの構造的に不可能であるキーワード一致以外の条件を用いた検索を行う．その方式として，Web ページの情報を RDF で記述し，それぞれのデータの関連付けと意味付けを行うことで，キーワード以外の条件での検索を実現し，その結果をユーザに推薦する．

3. 情報の推薦手法

本章では，1 章で述べた Web 検索における問題点と関連研究での問題点を解決するための具体的なアプローチを述べる．3.1 節では，現在のキーワード一致ではない，重要キーワードを用いた検索を実現するための Web ページ情報の記述に関する手法を述べる．3.2 節では 3.1 節で扱うデータの 1 つとして文章の重要キーワードの「特徴語」の抽出について述べる．3.3 節では，情報推薦に必要な検索情報を閲覧中の Web ページから取得する方法について述べる．

3.1 RDF を用いたページ情報の記述

現在の Web ページは HTML を用いて記述もしくは生成されることが多い．そのため，保持している情報がすべて文字列である．それに対し，本研究では RDF を用いて文章に記述されているトピック内の物事と Web ページの関連について，プロパティを用いて示す．それにより情報を検索する際の条件として利用する．関連は図 1 のように Web ページをリソースとして，そのページの重要な単語に対して「特徴語」というプロパティの付与や，その Web ページの重要な時期や新しさを関係付ける「時期」や「掲載日時」といったプロパティを付与する．「特徴語」については次節で詳細を説明する．RDF によって Web ページの情報を関連付けて記述することで，従来では文章内に埋まってしまっている関係を明示することができ，検索の際の条件にすることが出来る．また，RDF で記述することで様々な情報を自由に記述することも可能であり，ユーザが求める条件に合わせて Web ページに関連させる情報を自由に記述し，分野などによって独自の条件となる情報の追加を行う．様々な情報の例としては，カテゴリやジャンルなどの情報が挙げられる．

図 1 の例では，「土方歳三函館記念館」を説明した Web ページの情報を RDF で関連付ける例である．Web ページに対して，「土方歳三，函館戦争，海洋丸」が特徴語であることや，そのページが 8 月に重要な時期であること，2013 年 7 月 7 日が掲載日時でどのくらい新しい記事であるかなどを表現している．

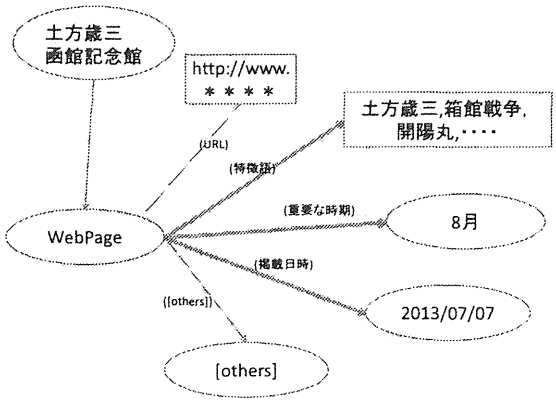

図 1 リソースの紹介を行う Web ページ情報関連付けの例
Figure1 example of Information to associate with the Web page

3.2 「特徴語」抽出

3.1 節で示した Web ページ情報の要素の 1 つとして，各 Web ページ内に含まれる文章を構成する重要キーワード(特徴語と呼ぶ)を定義する[5]．特徴語は，形態素解析処理で抽出される名詞から連続した名詞を結合することで形成され，分野固有の概念を含む専門性を持つ用語である．特徴語を用いて検索を行うことで，ユーザが着目している文章の分野を表す情報を取得することが可能になると考える．

文章は，Web ページ内の情報発信者が発信するトピックに関して説明した文章部分のことを指す．その文章を対象にキーワードの抽出と tf-idf(term frequency inverse document frequency)法を用いたキーワードのスコアリングを行い，本文中にある特徴語を抽出する．

tf-idf とは，式(3.1)で示される単語 i の文章 j における出現頻度 tf (term frequency) と，式(3.2)で示される単語 i を含む文章の数 df (document frequency), 及び文書の総数 N を用いて式(3.3)のように表わされる．これにより，文章 j における単語 i の特徴度合いを示す指標が得られる．

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3.1}$$

$$idf_i = \log \frac{N}{df_i} \tag{3.2}$$

$$tfidf = tf \cdot idf \tag{3.3}$$

この式によって得られる各単語のスコアを用いて，Web ページ内のトピックを説明した文章から特徴語を決定する．決定された特徴語は，「特徴語」のプロパティを付与し，RDF に記述する．

3.3 閲覧中の Web ページから検索キーワード抽出

次にユーザへリンクの推薦を行うための Web ページ内での情報抽出について説明する．3.1 節の Web ページ情報を持つ RDF に対する検索を行う際のユーザ側の検索トリガーを，ユーザの Web ページ閲覧とする．ユーザが閲覧するページ内の文章を抽出し，その文章から 3.2 節で取り上げた特徴語を検索要素として用いる．これは，トピックを説明する文章の中で，重要な単語をユーザが見落とさないためにシステム側が自動的に検索を行うために用いる．

4. 実装方式

3.1 節から 3.3 節で述べた推薦手法を実現するシステムの概要を図 2 に示し，以下で説明する．

(1) 閲覧中 Web ページの文章抽出

初めに，ユーザが現在閲覧中の Web ページの内容に興味を持ったと仮定する．この時，ブラウザではユーザが閲覧している文章を抽出するため，HTML の タグを利用して，文を抽出する．抽出された文は，結合して 1 つの文章として扱い，サーバに送信する．これらの処理は，Google Chrome ブラウザのアドオンである chrome extensions[6] を利用して実装する．

(2) 特徴語抽出

(1)でブラウザから送信された文章から，MeCab[7] による形態素解析を行い，抽出された名詞句を連結した「複合名詞」と呼ばれるキーワード群を生成する．キーワード群は 3.2 節で説明した tf-idf 法によってスコアリングを行う．MeCab の解析結果から複合名詞を生成する処理とスコアリング処理については，TermExtract[8] と呼ばれるモジュールを用いて処理を行う．

(3) 関連サイトの検索

(2)で得られたユーザが閲覧中の文章の重要語を用いて，同じキーワードが重要語として位置づけられているサイトを検索する．3.1 節で述べた Web ページの情報を持つ RDF に対して SPARQL で問い合わせを行い，Web サイトのリンクを取得する．

(4) 推薦リンクの提示

(3)の問い合わせによって得られた Web ページへのリンクをブラウザ上で表示する．図 3 では，ユーザが閲覧中の Web ページから抽出された文章を用いて特徴語を抽出する．その特徴語を用いて問い合わせを行い，取得される同じ特徴語を持つ Web ページのリンクが推薦されている．

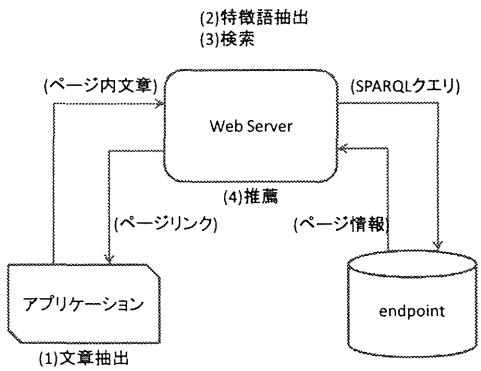

図 2 システム概要
Figure2 System summary

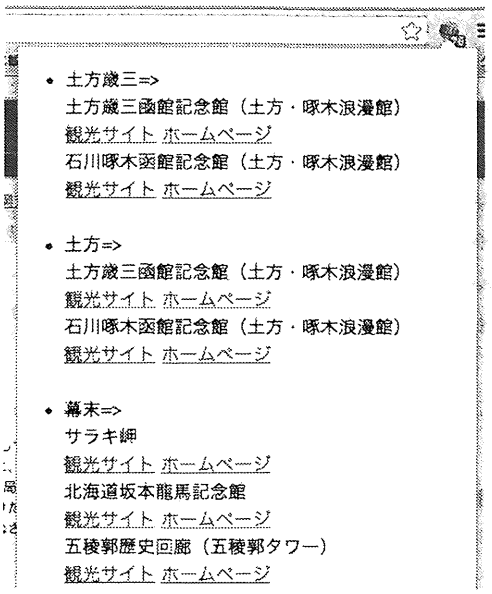

図 3 Web ページのリンク推薦

Figure 3 Recommendation of a Web page links

図 3 の例は，土方歳三の人物史を記述した Web ページの文章が抽出された際の推薦結果の例である．文章から抽出された「土方歳三，土方，幕末」という特徴語を用いて endpoint へ検索が行われている．その結果，リンク推薦として「土方歳三」に対しては「土方歳三函館記念館」などが推薦され，「幕末」に対しては「サラキ岬」などが推薦されている．

5. 評価実験

4 章で示した方式で実装を行い，アプリケーションを用いて評価実験を行う．5.1 章では，提案手法の有効性の範囲を定める調査について述べ，5.2 章では実験について述べる．

5.1 調査

提案手法が有効であるユーザやその状況を把握するため，別々な状況下で情報収集行うシナリオを用意し，それぞれの状況でユーザがどのような検索行動をとるのかを調査する．

5.1.1 調査概要

以下に示す 2 つのシナリオを利用して，被験者に検索を行わせる．

(1) **シナリオ 1**

"函館を何度か訪れたことがあるあなたは地元の友人と共に 8 月 10 日-11 日に道南観光を行います．あなたが友人と共に観光地を巡る計画を立ててください．"

(2) **シナリオ 2**

"函館を何度か訪れたことがあるあなたは 8 月 10 日-11 日に 1 人旅で道南観光を行います．どこを巡るか計画してください．"

推薦対象のデータとして，函館市観光サイト「はこぶら」[9]の一部データを RDF で記述する．推薦対象への検索条件として，3.1 節で述べた Web ページの要素のうち特徴語のみをクエリで扱っている．

シナリオ 1 では被験者 3 名(うち 1 名がアプリケーション不使用，2 名がアプリケーション使用)，シナリオ 2 では被験者 2 名(2 名ともアプリ使用)に対して道南の観光地に関する情報収集を行ってもらい，その中で 10 件訪問地を決定する作業を行なってもらう．アプリケーション調査開始前に説明を行い，任意での使用を求めた．訪問地は，行ったことがある，行ったことはないが知っている，初めて知った，の 3 つに分類する．

調査には，アプリケーションの利用を把握するためのログ取得と，ユーザがどのような場面でアプリケーションを利用するかを把握するための発話プロトコル法[10][11]を用いた実験を行った．

5.1.2 調査結果と考察

上記調査において，シナリオ 1 でのアプリケーションの有無によって行き先の分類に違いは見られなかった．次に，シナリオ 1 とシナリオ 2 では行き先の分類とアプリ有効性の違いが見られた．

表 1 調査結果

Table1 Results of the survey

	被験者	行ったことがある	行ったことはないが知っている	初めて知った	アプリ使用数	アプリ決定数
シナリオ 1	a	5	4	1		
	b	8	1	2	1	0
	c	4	5	1	1	0
シナリオ 2	d	3	2	5	1	0
	e	2	5	3	3	2

表 1 の要素はそれぞれ，被験者番号，「行ったことがある場所」を選んだ数，「行ったことはないが知っている場所」を選んだ数，「調査を行うまで知らなかった場所」を選んだ数，アプリケーションを使用した数，アプリケーションのリンクを使って決定した訪問地数を表す．

シナリオ 1 では，表 1 と決定したスポットから被験者本人が 1 回以上訪問したことがある有名スポット，もしくは行ったことはない函館の有名スポットを決定する傾向が強かった．これは，観光地調査において，その観光スポットが持つブランド価値が決定に影響を与えていると考えられる．被験者が閲覧している観光スポットの有名度や口コミから訪問地が決定され，ユーザはページの文章に着目しなかったためアプリケーションが使われなかったと考えられる．

シナリオ 2 ではシナリオ 1 と違い，知らない場所を選ぶ回数がシナリオ 1 よりも多いという傾向が見られた．また，知らない場所が決定されるプロセスにおいて，シナリオ 1 と違い，目的を持って調査を行う場合にはアプリケーションが有効である事例があった．目的を持った検索の例とし

て，”函館山に行きたい”という思考がある場合にはアプリケーションが使われず，”夜景を見たいから函館山に行く”と被験者が思考している場合，後者のばあいのみアプリケーションを利用して，ほかの夜景スポットを決定する事例があった．そこから，提案手法が目的を持って検索を行うユーザには有効であると考える．

5.2 実験

5.1 節の調査を踏まえて，アプリケーションの有効性を計測するために 2 つの実験を行う．

5.2.1 実験 1:目的の有無による情報収集の比較

調査で設定したシナリオ 2 に変更を加え，下記の条件を与えることで目的を持たせて被験者に情報収集を行わせた．

“函館・道南の＿＿＿＿＿＿＿＿について現地調査する．但し，上記テーマはいつ決めても構いません．また，すべてを無理矢理当てはめる必要もありません．”

上記の状況下での被験者の情報収集において，どのような行き先決定が行われるか，その中でアプリケーションがどのように利用されるか評価する．

5.2.2 実験 1 の結果と考察

実験 1 と調査のシナリオ 2 の被験者の利用にアプリケーションの利用回数に違いが見られた．表 2 の行き先分類の括弧内の数字は，決定数に対するアプリケーション利用からの決定数である．表 2 と図 4 からわかるように，情報収集の際に目的やテーマを持って調査を行うことで，被験者が興味を持ったページでのアプリケーション利用数がシナリオ 2 での利用数よりも多くなることが分かる．また，アプリケーションで推薦されたリンク先のページを閲覧した時の行先決定数も同様に占める割合が多くなったことが分かる．これらの結果より，目的を持った情報収集を行った場合に，ユーザが着目した Web ページにある特徴語で検索を行った結果を提示することは，ユーザの情報収集には有効であると考えられる．

表 2 実験 1 結果

Table2 Results of the research Ⅰ

被験者	行ったことがある	行ったことはないが知っている	初めて知った	アプリ使用数	アプリ決定数
A	4	2	4(1)	5	1
B	0	3(1)	5(3)	16	4
C	0	3(2)	7(2)	9	4
D	5(2)	4(3)	1	10	5
E	3	2	2(1)	4	1

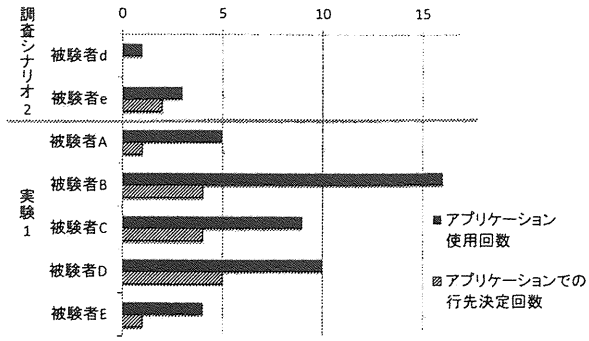

図 4 アプリケーション利用の比較

Figure4 Comparison of application usage

5.2.3 実験 2:閲覧 Web ページとの比較

提案方式の有効性と Web ページに貼り付けられているリンクとの優位性を確かめるため，ユーザが閲覧中の Web ページ内にあるリンクとアプリケーションが推薦するリンクのどちらが有用な情報であるかを検証するための比較実験を行う．実験 1 と同様の状況下での調査において，被験者が興味を持ったページで義務的にアプリケーションを利用してもらう．その上で，閲覧中の Web ページに存在するリンクや関連記事に着目したのか，アプリケーションから推薦されるリンクに着目したのかを比較する．

5.2.4 実験 2 の結果と考察

表 3 の被験者の行き先の分類に着目すると，義務的にアプリケーションを使用してもらい，閲覧中の Web ページに存在するリンクよりもアプリケーションから推薦したリンクをたどって「初めて知った」観光スポットを行先として決定する割合が全体のうちの多くを占めるユーザが多くみられた．さらに，図 5 のグラフで示すように「初めて知った」観光スポットを決定する際に，アプリケーションリンクからの推薦で決定される割合が多くを占めた．

これは，カテゴリツリーなどに初めて知る情報が表示される場合には，被験者はあまり有用な情報であると考えなかったが，被験者が着目したページに関連付けて初めてしる情報を提示された場合には，目的に沿った情報が得られると期待したためであると考えられる．また，これら結果から，着目したページにはユーザの目的を満たすリンクがあまり貼られていなかったこと，ページの文章に含まれる特徴語と同じ特徴語を持つページのリンクを推薦することは有効であることが分かった．

表 3 実験 2 結果

Table3 Results of the research Ⅱ

被験者	行ったことがある	行ったことはないが知っている	初めて知った	アプリ使用数	アプリ決定数
F	3(2)	1(1)	6(5)	12	8
G	3(1)	2(1)	5(3)	11	6
H	0	3(1)	7(5)	12	6
I	0	4(2)	6(6)	9	8
J	6	1(1)	3	12	1

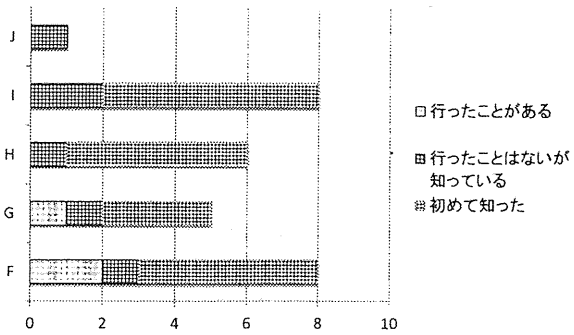

図 5 アプリケーションによる行き先決定数と内訳

Figure5 Breakdown destination and number of decisions by the application

6. おわりに

本稿では，現在の Web 検索ではユーザの目的とした情報が必ずしも取得できない問題を解決する情報推薦手法の提案を目標として，RDF を用いた Web ページ内情報と Web ページに関する情報の構造化と文章の内容を示す特徴語を用いた情報検索と推薦手法を提案した．提案した手法を実装し，函館の観光サイトの情報を対象に推薦実験を行ったところ，目的を持って情報収集を行うユーザには提案手法が有効であることが確認できた．

今後は，提案した Web ページ情報の RDF 化のうち，時期や投稿日時等扱わなかった情報を踏まえた推薦や，tf-idf 法を用いて特徴語のスコアリングを行った際の値を用いた特徴語の重要度などについて考慮していく．さらに，今回の実験で推薦対象とした観光分野以外の情報を扱うことで本稿の提案が汎用的な情報に有効であることを検証していく．

参考文献

1) Christian,Tom,Tim, Linked Data の仕組み，情報処理学会誌，Vol.52, No.3, pp.284-292, 2011.
2) 渡辺奈夕子，岡本昌之，菊池匡晃，飯田貴之，佐々木健太，堀内健介，山崎智弘，大村寿美，服部正典，閲覧 Web ページからの第 1 検索キーワード抽出に基づく検索支援，情報処理学会論文誌,Vol.53,No.7,pp1783−1796,2012.
3) Resource Description Framework (RDF),http://www.w3.org/RDF/, 2004.
4) 大西可奈子，小林一郎，Linked Data から得られるリソース間関係に着目した情報拡張手法の提案，情報処理学会研究報告，Vol2011-IFAT-No.4，2011.
5) 中川裕志，森辰則，湯本紘彰，"出現頻度と連接頻度に基づく専門用語抽出"，自然言語処理，Vol.10 No.1, pp27-45, 2003.
6) chrome.extension, http://developer.chrome.com/extensions/extension.html, 2013.
7) mecab -Japanese morphological analyzer-, https://code.google.com/p/mecab/, 2013.
8) 専門用語(キーワード)自動抽出用 Perl モジュール"TermExtract"の解説，http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html
9) 函館市公式観光情報サイトはこぶら，http://www.hakobura.jp/.
10) 海保 博之，原田 悦子，プロトコル分析入門—発話データから何を読むか，新曜社，1993.
11) 三輪眞木子，江草由佳，齊藤ひとみ，髙久雅生，寺井仁，神門典子，Web 上の exploratory search の特徴:発話プロトコルと事後インタビュー分析結果より，情報処理学会研究報告．情報学基礎研究会報告 2009-FI-96(2), 1-8, 2009-11-12.