

# ユーザの Web 探索履歴における キーワード遷移に基づく Web ページ推薦システム

枝 隼也<sup>†</sup> 佐藤 哲司<sup>††</sup>

<sup>†</sup> 筑波大学図書館情報専門学群 〒305-8550 茨城県つくば市春日 1-2

<sup>††</sup> 筑波大学図書館情報メディア研究科 〒305-8550 茨城県つくば市春日 1-2

E-mail: <sup>†</sup>s0813167@u.tsukuba.ac.jp, <sup>††</sup>satoh@slis.tsukuba.ac.jp

あらまし 近年 Web ページの数が急激に増加するのに伴って、ユーザは様々なページを閲覧・比較するなどの探索行動を繰り返して所望のページにたどり付かなければならない場合が増えている。本論文では、ユーザが閲覧する Web ページのキーワード間の遷移に基づいて Web ページを推薦する手法を提案する。連続して閲覧するページ間でキーワードをノードとする有向グラフを生成し、キーワード間の連結強度からキーワード遷移を推定する。そして、推定したキーワードを用いてユーザが閲覧するであろうページを推薦する。提案法を実装したシステムで抽出したキーワードを使用した主観評価を行ない、Web 閲覧行動のパターンを推定できる場合があること、および、キーワード遷移グラフで推定したキーワードが有効に機能することを確認した。

キーワード 閲覧履歴、共起度、グラフ、推薦

## Web-page Recommendation System based on the Keyword transitions through a Web Exploration

Junya EDA<sup>†</sup> and Tetsuji SATOH<sup>††</sup>

<sup>†</sup> School of Library and Information Science, University of Tsukuba

1-2, Kasuga, Tsukuba, Ibaraki, 305-0855 Japan

<sup>††</sup> Graduate School of Library Information and Media Studies, University of Tsukuba

1-2, Kasuga, Tsukuba, Ibaraki, 305-0855 Japan

E-mail: <sup>†</sup>s0813167@u.tsukuba.ac.jp, <sup>††</sup>satoh@slis.tsukuba.ac.jp

**Abstract** In recent years, The number of web-pages are increasing considerably. User cannot find a desired page without much burdens. We propose an efficient method to recommend web pages, which user want to visit. We create the directed graph which node are keyword between pages and estimate keyword transition from connection strength between the keyword. We recommend the Web-pages that a user will read with the keyword which we estimated. We evaluated that we used the keyword which I extracted by the system. we could estimate the pattern of the Web reading action and confirmed that the keyword which I estimated by a keyword transition graph functioned effectively.

**Key words** Web exploration logs, co-occurrence frequency, recommendation

### 1. はじめに

近年 Web 空間のページ数は急激に増加し続けている。Web には誰でも情報を記述することができるため、次々と新たなページが生まれる一方、古いページも残ったまま、重複した内容のページが様々な場所に点在している。また、著者が匿名であるページや、内容の信憑性が疑わしいページも数多く存在する。このような特徴から、ユーザは興味や関心に応じて何度も検

索を繰り返し多くのページを閲覧して比較検討して必要な情報を得ているのが現状である。このような様々なページを辿るという探索行動は、所望のページにたどりつくまでの手間を増やし、ユーザの負担となっている。この問題を解決するために、ユーザの情報探索における興味を的確に捉えて情報推薦を行おうとする様々な研究がなされている。しかしその研究の多くは、URL やページの移動順に注目したものであり、個々のページの内容まで利用し推薦を行おうとする研究は緒についたばかり

である．また，ユーザの Web 閲覧と探索行動における興味の変化を随時反映し推薦に利用しようとするものもほとんど知られていない．

本論文では，ユーザが閲覧するページを移動することで変化する話題からユーザの興味を推定する手法を提案する．ユーザが閲覧したページから話題を構成するキーワードを抽出し，ページの移動によって変化する話題のキーワード遷移に基づいてユーザの興味を推定する．ここでキーワード遷移とは，各ページに出現するそのページの話題を表すキーワードをノード，ページ間にまたがるキーワード間の結合係数をエッジの重みとした非循環有向グラフとする．ユーザがページを移動するたびにキーワード遷移グラフを再計算し，ユーザの興味を推定し興味に合った文書を推薦する．このように提案法は，ユーザの Web 探索履歴を用いて Web 文書の推薦を行い，URL やページの移動順の他にそのページの内容まで利用することに特徴がある．また，ユーザの Web 探索行動の過程で変化する興味を推薦に随時反映することも特徴である．

以下 2 章で，ユーザへの情報推薦に関する関連研究について述べ，本研究の位置づけを示す．3 章で，本研究が提案する，キーワードの遷移に基づく Web ページの推薦手法を説明し，4 章でその手法を実装したシステムの詳細について説明する．5 章で，評価実験について説明し，考察を行った後，6 章でまとめと今後の課題について述べる．

## 2. 関連研究

ページの内容に基づく推薦とユーザの Web 閲覧履歴に基づく推薦に分けて関連研究を概観し，本研究の位置づけを明確にする．

### 2.1 ページ内容に基づく推薦

文書内に出現する単語の統計情報を用いて推薦を行う研究が盛に行われている．またページに出現する，単語の類似度からページをクラスタリングする手法も知られている．

ユーザの興味に沿った広告の推薦を目的として，複数のテーマが混在した文書からその文書内容をよりの確に表したキーワードを抽出する研究に Grineva [1] がある．文書内の単語をノードとし，Wikipedia を辞書として意味のつながりをネットワーク構造で表す．文書内の単語の関連性を明らかにし，複数のトピックが混在する文書から内容を表す特徴的なキーワードを抽出する手法を提案している．

陳ら [2] は，Web コンテンツ間の関連の強さを類似度で表現し，ユーザの検索時に個々のリンク距離に類似度を反映することで，ユーザが指定したページからリンクを辿って形成する探索空間に関連性の高いコンテンツをより多く含ませる手法を提案している．また鶴原ら [3] は独立成分分析を用いて，文書のベクトル空間からトピックと呼ばれる特徴軸を見つけ出し，その特徴軸を用いて，似た軸を持つページを推薦する手法を提案している．

### 2.2 ユーザの閲覧履歴に基づく推薦

個人の閲覧履歴等を用いて，検索を支援するパーソナライズと，他のユーザの履歴を組み合わせる支援を行う協調フィルタ

リングとがある．

パーソナライズの研究例には，ユーザの Web 閲覧履歴に出現する単語をクラスタリングし，その結果をユーザプロファイルとしてマッピングし検索語に拡張利用する手法 [4] が知られている．また，あるサイトのページ間のユーザの移動をモデル化することで，サイト内の閲覧をスムーズなものにしようとする行動ターゲティングと呼ばれる手法も知られている．行動ターゲティングとは，閲覧，回遊状況などの web 上のユーザの行動履歴に応じて，配信するコンテンツをパーソナライズする技術である．山本ら [5] はユーザの行動履歴に基づいたページ間ネットワークを作成し，そこから得られる様々な属性を用いた行動ターゲティング手法を提案している．

他のユーザのページ移動など他人の閲覧履歴を用いて，そのユーザとよく似た行動をする別のユーザにページを推薦する手法を協調フィルタリングと言い，代表的な研究例に岩田ら [6] がある．同じものを何度も推薦してしまう協調フィルタリングにおける問題点への対策として，オンラインストアにおける商品の購買順序を考慮した確率モデルを導入することで，従来よりも高速で予測精度が高い推薦が行えるとしている．

### 2.3 本研究の位置づけ

ユーザの閲覧履歴を用いた研究は，URL やページの移動順などに注目したものが多く，そのページの内容まで利用し推薦を行おうとするものは知られていない．また，ユーザの閲覧行動における興味の変化を推薦に利用しようとする研究もほとんど知られていない．

本論文では，ユーザの閲覧履歴情報とページの内容の解析を組み合わせる推薦を行う手法を提案する．さらにグラフ構造を用いてキーワードの時間的な推移を追うことで，ユーザが連続してページを閲覧する際の興味の変化に対応したキーワードやページの推薦提示を行う．

## 3. キーワード遷移に基づく提案法

### 3.1 提案法の概要

提案するキーワード遷移に基づくページ推薦処理の流れを図 1 に示す．提案法では，ユーザが閲覧した Web ページの URL を入力とし，そのページの内容を表すキーワードを抽出する．次に，各ページ間でキーワード遷移のグラフを構築する．グラフ構造を解析して，ユーザの興味を表すキーワードの組み合わせを選択し，そのキーワードに関する内容のページを推薦する．以下，この手順にしたがって，各処理を詳細に述べる．

### 3.2 Web ページ間キーワード遷移

本論文では各ページに出現するそのページの話題を表すキーワードをノード，キーワード間の *Simpson* 係数の値をエッジの重みとした無向グラフをキーワード遷移グラフという．グラフのノードとなるキーワードは Web ページから抽出する．なお，キーワード抽出の詳細は 4.3 節で詳しく処理を説明する．

事象の関連を確率的な過程として，有向グラフを用いて表す方法がベイジアンネットワークである．このネットワークは閉路を持たない有向グラフ，非循環有向グラフ (directed acyclic graph) でなければならない．有向とすることで，事象間の影響

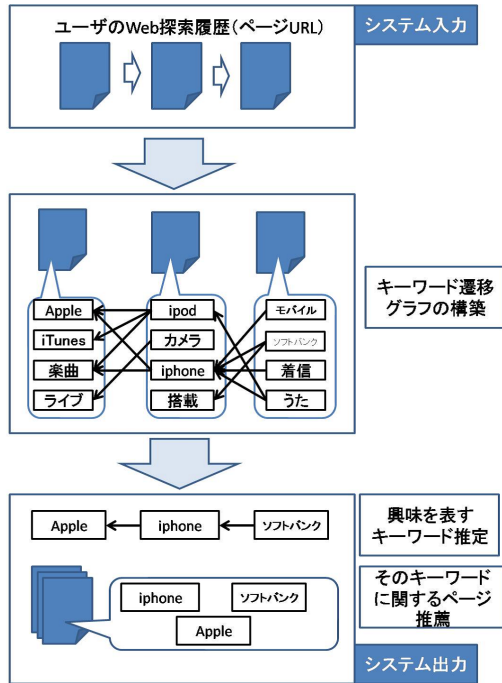


図 1 提案手法の概要

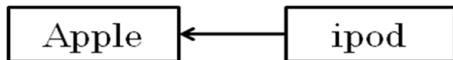


図 2 ノードとエッジの例

(因果関係)の方向性を表すことができる。ユーザの閲覧履歴に適用するにあたり、ユーザが閲覧している最新のページがそのユーザの興味を最も良く表していると考え、そこで、最新ページから古いものへと順次接続する有向グラフを考える。また、グラフのエッジは、ノードとなるキーワード間の *Simpson* 係数でスコアを与えることとする。ここで非循環とは、ネットワークの有向辺をたどって移動したとき、移動の経路上に帰ることがない、すなわち、原因と結果が循環的な構造をなさないことである。

非循環有向グラフでキーワード遷移を表現することで、あるページのキーワードAにつながるキーワードは決してAに戻ることはなく、より新しいページのキーワードを起点として、より古いページのキーワードに向かうキーワードの連鎖を、複数のパスとして表すことができる。このようなキーワード間の依存関係を表した非循環有向グラフを、本論文ではキーワード遷移グラフという。

### 3.3 キーワード遷移からの興味推定

#### 3.3.1 ノードとエッジの決定法

キーワード遷移グラフのノードはキーワードであり、エッジの重みは両端のノードキーワードの共起度とする。キーワードの抽出には形態素解析 *mecab*<sup>(注1)</sup>と専門用語抽出システム *Termextract* [7] を用いた。さらに、共起度の計算法については次に記す *Simpson* 係数を用いる。

ノードが「Apple」と「ipod」からなるキーワード遷移グラ

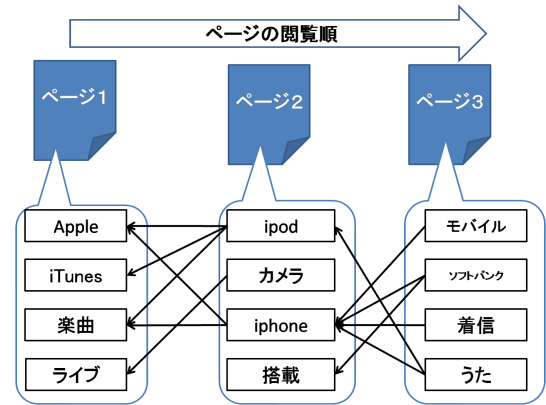


図 3 キーワード遷移グラフの例

フの例を図 2 に示す、この例では、 $Simpson(Apple, ipod)$  がエッジの重みとなる。ここで、*Simpson* 係数とは、単語間の共起頻度に基づく値である。単語  $X$  と単語  $Y$  が単独で出現する文書数をそれぞれ  $|X|$  と  $|Y|$ 、 $X, Y$  の 2 つの単語が同時に出現する文書数を  $|X \cap Y|$  とすると、*Simpson* 係数は、

$$Simpson(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)} \quad (1)$$

で与えられる。すなわち、分母は各語が含まれる文書数の小さい方、分子は 2 つの語がともに含まれている文書数とした比である。分母に各語の出現文書数の少ない方を用いることで、出現数が少ない語から見た、語の共起度を計算することができ、各単語の出現文書数に極端な差がある場合でも有用な計算結果を出すことができるとされている。

本論文で示すキーワード遷移グラフのエッジの重みは、*Simpson* 係数を拡張した松尾ら [8] の閾値付き *Simpson* 係数 (2) とした。閾値を設けることで、単独ヒット件数が極端に少ない語の係数が正確に計算できなくなってしまう問題を防ぐことができる。

$$Simpson(X, Y) = \begin{cases} \frac{|X \cap Y|}{\min(|X|, |Y|)} & \text{if } |X| > k \text{ and } |Y| > k, \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

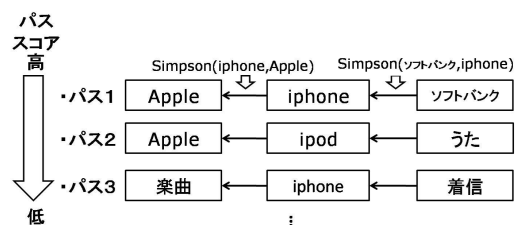


図 4 パスの選択方法

(注1): <http://mecab.sourceforge.net/>

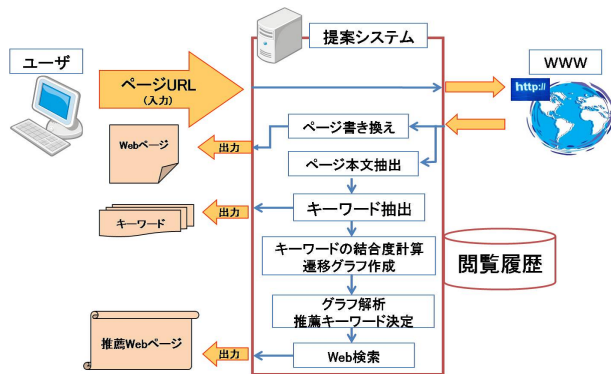


図 5 システム処理概要

### 3.3.2 パスのスコア計算法

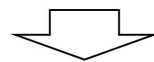
現時点でユーザが閲覧しているページから古い方向へエッジのスコアの対数を加算していく。そのスコアが最大となるパスに出現する語を検索クエリとし、それをを用いて検索エンジンによって検索された Web ページを推薦結果とする。図 3 は、システム内で構築されているグラフ構造を実際の例を用いて表したものである。ページ 3 が現在閲覧している最新ページで、ページ 2 は一つ前に閲覧したページ、ページ 1 が二つ前に閲覧したページである。縦に並ぶ 4 つの単語がそれぞれのページ内で抽出された特徴語を表している。閲覧順に閾値を超える Simpson 係数を持つ語同士をエッジで繋ぐことによりキーワード遷移グラフを作成する。こうして構築されたグラフ構造より、各エッジの Simpson 係数を足し合わせすべてのパスのスコアを計算する。モデルで示した計算法を具体例を用いて説明する。まず、最新文書であるページ 3 から抽出されたキーワード「モバイル」に着目する。ノード「モバイル」からは、ページ 2 のノード「iphone」へエッジが通っている。その「モバイル」-「iphone」間のエッジのスコアは  $Simpson(モバイル, iphone)$  である。更に、ノード「iphone」からページ 1 のノード「Apple」へパスがつながっている。そのスコアは  $Simpson(iphone, Apple)$  であるので、「モバイル」から「iphone」を通り「Apple」へ至るパスのスコアは、 $\log \{Simpson(モバイル, iphone)\} + \log \{Simpson(iphone, Apple)\}$  となる。

図 4 は計算されたスコアを高い順に並べた例である。最新ページのページ 3 から一番古いページであるページ 1 にたどり着く全てのパスについて、そのパスが持つ全エッジの値を加算することで、パスのスコアとする。得られたスコアを降順にソートし、最も高いスコアのパスを構成しているキーワードがユーザの興味を表している語とする。その語を内容として持つページを推薦する。ページ推薦に関する詳細な処理は次章で説明する。

## 4. Web 文書推薦システムの実装

提案法を実装したシステムの処理の概要を図 5 に示す。ユーザは本システムを通して Web 閲覧を行う。本システムはユーザが閲覧するページの URL を入力とし、「閲覧する Web ページ」

```
<a href=
"http://japan.cnet.com/tag/iPad/"
title="アップル「iPad」は旋風を起こす?" target="_self">
アップル「iPad」は旋風を起こす?</a>
```



```
<a href=
"http://ce-lab/system.php?http://japan.cnet.com/tag/iPad/"
title="アップル「iPad」は旋風を起こす?" target="_self">
アップル「iPad」は旋風を起こす?</a>
```

図 6 ページ書き換え処理

「そのページの内容を表すキーワード」「ユーザの興味を反映した推薦ページ」をユーザに提示する。システムに実装した、図 5 に示す 5 つの機能について、順次詳細に説明する。

### 4.1 ページ書き換え

ページ書き換えの具体的な処理を図 6 に示す。本システムでは、閲覧中の Web ページの URL を受け取り履歴として保存する。ユーザのリクエストしたページの URL を受け取り、次のリクエストも本システムを経由するように、ページ内のリンクをすべて書き換える。全てのリンクを書き換えたページをユーザに提示するが、書き換えはリンクだけなのでユーザの閲覧に支障はない。閲覧中のページに含まれる全てのハイパーリンクをページ移動のたびに書き換えることで、リンクを辿り複数のページを閲覧する連続した Web 閲覧行動にも対応できる。

### 4.2 ページ本文抽出

Web ページには HTML のレイアウト情報や広告などコンテンツの本文とは直接関係していない情報も多く含まれている。ページの主題と思われるテキストを処理の対象とするには、記事の本文と思われるテキストだけを抽出する処理が必要となる。

ここではまず Web ページから HTML タグなどの情報を除いた本文テキストを抽出するモジュール Contentextract<sup>(注2)</sup>について説明する。まず、ページを記述している HTML 文書から不要な HTML タグを削除する。その際にタグ「div」「td」で囲まれた範囲をブロックとして分割する。各ブロックに、句読点の数が多くほどスコアが高くなるように、リンクタグが多いほどスコアが低くなるように本文らしさスコアを付与する。本文らしさスコアが高いブロックをつなげてクラスタとし、スコアの一番高いクラスタを本文とする。以上の処理によって Web ページからタグやコンテンツと直接関係のない情報を取り除いた本文を抽出する。

### 4.3 キーワード抽出

本文テキストからキーワードを抽出し、重要度順にスコア付けを行う。Contentextract によって抽出された Web ページの本文を形態素解析し、その結果に Termextract [7] を適用する。この処理によって、本文テキストに対して順位づけられたキーワードのリストが得られる。

Web ページからキーワードを抽出した例を図 7 に示す。この図は CNET Japan<sup>(注3)</sup> 内の「MOBILE CHANNEL」記事の

(注2): <http://www.systemfriend.co.jp/node/326>

(注3): 出典:朝日インタラクティブ CNET Japan <http://japan.cnet.com/>



図 7 ページからのキーワード抽出例

Web ページから、キーワード 10 個を抽出している。図の左側が記事のページ<sup>(注4)</sup>、右のリストが抽出された「Google」「携帯電話」などのキーワードである。

#### 4.4 グラフ構造による表現と解析

抽出されたキーワードを、3.3 節で示した方法によりグラフ構造に表現する。語をユーザの移動元と移動先でペアにし、その語同士の結合度を *Simpson* 係数によって計算する。この計算には Yahoo!WebsearchAPI<sup>(注5)</sup>を用いて各単語をクエリとして検索したヒット件数をその語の出現する文書数とした。さらに 2 つのキーワードを AND 条件でつないだクエリのヒット件数をその 2 つのキーワードがともに出現する文書数とした。Web 検索のヒット件数を用いることにより、事前に膨大な数の単語の出現文書数を算出しておく必要がなく、新たな語、未知語にも柔軟に対応することができる。ある単語の出現する文書数を  $m, n$  とすると、 $Simpson(m, n)$  で単語同士の結合度を表す。

一定の値以上の語のペアをエッジでつなぐことによって、新しいページから古いページ方向に向く非循環有向グラフ構造を表した。この処理によって、順位づけられたキーワードのリストから、重み付きのグラフ構造が作られる。

提案法を実装したシステムでは *Simpson* 係数の値が  $0.2 \leq X \leq 0.8$  でエッジを繋ぐこととした。参考文献 [8] に示された *Simpson* 係数の下限 0.2 を採用することとした。また、0.8 以上となる語のペアは、同じ意味を表す異表記、たとえば「Google」と「グーグル」などや、どの文書にでも出現する一般的な語が多かったためエッジを繋がない事とした。また、閾値付き *simpson* 係数 (2) の閾値  $k$  は、3000 とした。単独ヒット件数が 3000 件以下である語は、キーワード抽出の誤りによる不自然な語であることが多かったためである。

#### 4.5 Web ページの推薦

はページの推薦処理の流れを図 8 に示す。3.3.2 の手法によって解析されたグラフから、キーワードの組み合わせを取得し、その組を Yahoo!WebserchAPI に渡し、検索にヒットしたページをシステムが提示する推薦ページとする。この処理によって単語のグラフ構造から最終的な出力 Web ページの URL が得

(注4): この記事は海外 CBS Interactive 発の記事を朝日インタラクティブが日本向けに編集したものです。

<http://japan.cnet.com/mobile/story/0,3800078151,20406230,00.htm>

(注5): <http://developer.yahoo.co.jp/>

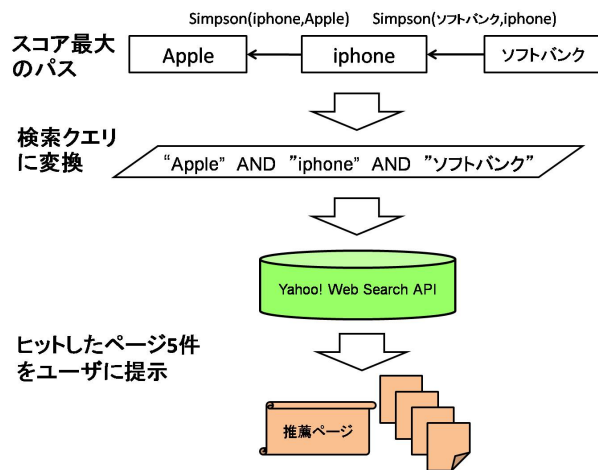


図 8 ページ推薦処理

られる。

試作した Web システムの画面イメージを図 9 に示す。画面の左フレームに現在閲覧中のページの特徴語、中央のフレームに閲覧中のページ<sup>(注6)</sup>、右フレームに推薦結果のページへのリンクを 10 件表示している。

## 5. 評価実験

### 5.1 実験概要

前項 4.3 で示したように Web ページから話題語が抽出されることから、この抽出された話題語がユーザの興味を適切に表現しているかを、利用者実験によって確認した。実験を行うにあたり、一般的なユーザの連続する Web 閲覧行動は以下の 4 つのパターンに分類できるとした。

- A. 同じ話題の別のページを見ていくもの
- B. 広い話題から徐々に狭い話題に絞り込んでいくもの
- C. 狭い話題から徐々に広げていくもの
- D. ある話題を徐々にシフトしながら追っていくもの

この 4 パターンでユーザの閲覧行動が全て言い表せている保証はないが、利用者実験によって少なくともこの 4 パターンの違いが明らかになるかを調査することは意義があることだと考える。実験参加者が正しく分類することができれば、ユーザの Web 探索行動の特徴を、キーワードのリストによって正しく捉えることができていくといえる。

### 5.2 実験に用いたデータと質問

CNET Japan 内の記事を各分類パターンに沿って、実際に閲覧した探索履歴のデータを用いて実験を行った。表 1 から表 3 は CNET Japan 内の記事を 3 つ続けて閲覧したページのキーワードの履歴である。

表 1 は前項 5.1 で示した閲覧行動の分類パターンのうち「D. ある話題を徐々にシフトしながら追っていくもの」に当てはまる

(注6): <http://japan.cnet.com/news/media/story/0,2000056023,20394669,00.htm>





図 9 システム動作の様子

閲覧履歴である。同じように表 2 は「B. 広い話題から徐々に狭い話題に絞り込んでいくもの」、表 3 は「C. 狭い話題から徐々に広げていくもの」のパターンに当てはまる閲覧履歴である。

ページ 1 が現在閲覧しているもの、ページ 2 はひとつ前に閲覧したページ、ページ 3 は二つ前に閲覧したページである。

問 1. Web 閲覧行動を表現するキーワードリスト評価  
 < 閲覧パターン分類 >  
 この 3 つのページを閲覧したユーザはどのような興味を持っていたと考えられますか。さらに、この閲覧は前項 5.1 の閲覧行動の分類で示したどのパターンに当てはまると思いますか。

問 1 ではキーワードリストでユーザの Web 閲覧行動を表しているかを評価するため、ユーザに提示されるキーワードの遷移を示したリストから、そのリストはどのような探索行動を表わすかを問う、閲覧パターン分類実験を行った。さらに、実験を行うにあたり分類した Web 閲覧行動の 4 つのパターンのどれに、提示したリストの閲覧履歴が当てはまるか回答してもらった。

問 2. Simpson 係数を用いたキーワード遷移グラフ評価  
 < 各文書からキーワード選択 >  
 同じ興味で閲覧を続けていく場合、この次のページに移るにはどのような検索語が適当であると考えられるか。各ページから優先度をつけてキーワードを 3 つずつ選べ。

問 2 では、Simpson 係数を用いたキーワード遷移のグラフの評価として、同じく提示したキーワードのリストからページの推薦に有効だと思われるキーワードを選択してもらった。

履歴データ 3 つに対し、情報系を専攻する大学生 6 人に以上の問いに答えてもらった。

### 5.3 結果

#### 5.3.1 閲覧パターン分類の結果

履歴 1 から 3 に対する閲覧パターン分類 (問 1) の回答を表 4 に示す。履歴 1 は履歴行動パターン D、履歴 2 はパターン B、

表 1 履歴データ 1

ページ 1	ページ 2	ページ 3
Bing	Microsoft	検索
検索	検索エンジン	Bing
Microsoft	Bing	シェア
検索エンジン	シェア	Microsoft
利用浸透	検索市場シェア	機能
検索者	Hurt	検索エンジン
宣伝費投入	検索	Google
宣伝	滑り出し	Wolfram Alpha
テクノロジー	comScore	情報
Bing Cashback	増加	検索市場

表 2 履歴データ 2

ページ 1	ページ 2	ページ 3
無線 LAN	機種	冬春モデル
Mbps	キー部分	キャリア
携帯電話	位置情報	ラインナップ
NTT ドコモ	冬春モデル	正式発表
最大	端末	春モデル
携帯型ゲーム機	ディスプレイ	ニュース
カメラ	セパレートケータイ	冬モデル
印刷用ページ	Android 端末	国際家電見本市
デジタルカメラ機能	docomo PRO series	Nexus One
瞬速起動	docomo STYLE series	NTT ドコモ

履歴 3 はパターン C となるように作成したデータである。表 4 の結果から、実験者は提示されたキーワードリストを見て、履歴行動パターンをある程度判別できていることがわかる。履歴パターン A と B、および履歴パターン C と D の判別は難しいが、A,B と C,D を取り違える実験者はいなかった。

#### 5.3.2 各文書からキーワード選択

実験者が問 2 で回答した語を、回答した優先度に応じて 3、

表 3 履歴データ 3

ページ 1	ページ 2	ページ 3
スマートフォン	iPhone	Nexus One
Forrester	Google	Android
iPhone	Google ケータイ	搭載携帯
携帯電話市場	ビデオ撮影	携帯電話
年	動画撮影機能	正式発表
iPhone OS	アップル	HTC
BlackBerry	GS	グーグル
BlackBerry OS	機能 OS	Google
スマートフォン向け OS	アプリ販売サイト	下
モバイル OS メーカー	Android マーケット	ディスプレイ

表 4 問 1. 閲覧パターン分類の結果

	履歴 1	履歴 2	履歴 3
U1	A	B	D
U2	A	B	C
U3	B	A	D
U4	A	B	C
U5	B	C	D
U6	A	B	D

表 5 履歴データ 1 においてシステムが提示した語

ページ 1	ページ 2	ページ 3
提示した語	提示した語	提示した語
Bing	検索市場シェア	Google

2, 1 点を加算して集計した結果を表 8 に示す。

実験者は各文書で共通して登場する語、何度も登場する語を高い優先度で選んでいる。同じ履歴データに対して提案システムが提示するキーワードの組み合わせを表 5 に示すのである。提案システムは、各ページから *Simpson* 係数によって結びつけられたキーワードうちの一番スコアが大きい組み合わせを提示する。システムが提示した語のうち「Bing」「検索市場シェア」という 2 つの語に関しては、実験によってユーザも高い優先度を与えている。しかし、システムがページ 3 から抽出した「Google」という語は実験者には全く選ばれなかった。このことは、時系列において実験者が見た新しいページから 2 つ目まではシステムが抽出したキーワードと実験者の選択が一致したこともある。

表 6 履歴データ 2 においてシステムが提示した語

ページ 1	ページ 2	ページ 3
提示した語	提示した語	提示した語
NTT ドコモ	セパレートケータイ	ニュース

次に履歴データ 2 におけるパターン B を選択した実験者の各文書からキーワード選択 (問 2) の回答を表 9 に示す。パターン B を選択した実験者は、「無線 LAN」「docomoPROseries」など特定性の機能や商品を表す語を選んでいる。これは話題を狭めていくために、より特定の事柄を示す語を選んでいこうとしているためだと考えられる。同じ履歴データにおいて、システ

表 7 履歴データ 3 においてシステムが提示した語

ページ 1	ページ 2	ページ 3
提示した語	提示した語	提示した語
iPhone OS	Google	搭載携帯

ムが提示するキーワードの組み合わせを表 6 に示す。履歴データ 1 の結果と同様に、新しい文書から 2 つ目までの語、「NTT ドコモ」「セパレートケータイ」は実験でユーザが選んだものと重なっている。しかし、一番古い文書であるページ 3 からシステムが提示した「ニュース」という語は、全ての実験者が選択していなかった。

次に履歴データ 3 におけるパターン D の人の問 2. 各文書からキーワード選択の回答を表 10 に示す。同じ履歴データにおいて、システムが提示するキーワードの組み合わせを表 7 に示す。履歴データ 3 の結果では、一番新しい文書の語、「iPhone OS」実験で一人のユーザが選んでいる。しかし、その他の文書からシステムが提示した「Google」「搭載携帯」という語は、全ての実験者が選ばなかった。

#### 5.4 考 察

表 4 の結果から、実験者は提示されたキーワードリストを見て、履歴行動パターンをある程度判別できていることがわかり、キーワードリストによって Web 探索行動の特徴を表現できているといえ、ユーザの Web 探索行動をキーワードリストの積み重ねで表現することの有効性が示されたと考えられる。

表 8, 表 9, 表 10 より実際にシステムによって抽出されたキーワードと被験者によって選び出されたキーワードを比べてみると、システムが選んだものと同じものもいくつか選ばれ、更に似た内容を表す語が優先度が高くユーザに選ばれていることから、*Simpson* 係数によってキーワードの遷移を表すことの有効性が示された。しかし、システムが選ぶ語の組み合わせと、全く同じものを選ぶユーザがいなかったこと、最新ページから 2 つ前までしか有用だと思われるキーワードが抽出できていなかったことなどの結果より、Web の閲覧行動のパターンや特徴に応じて、そのパターンを考慮したシステムのキーワード選択の手法に拡張する必要があると考えられた。

また、本手法を用いたシステムでは、「Bing」「セパレートケータイ」「iPhoneOS」などかなり限定的な内容を示すキーワードが選ばれていたことから、話題を狭めていくようなパターンの Web 探索行動には特に有効なのではないかと考えられた。

Web 探索の途中でそのユーザの興味が全く違うものになってしまった場合、本システムでは有効な推薦は行うことが出来ない。現段階では、システム内の履歴を削除する、というボタンをユーザが押すことで新たな興味での推薦に対応している。しかし、ユーザが自ら申告せずともシステムが興味の変化を抽出することが理想であるため、そのような処理も行えるようにするといった事が課題にあげられる。具体的には、ページ同士のグラフ構造を作成する際のエッジの数などで興味の切れ目が推測できるのではないかと考えられる。

ユーザの履歴管理に関して、現時点ではページ書き換えによって URL を入手し保存しているが、プロキシサーバを用い

表 8 履歴データ 1 におけるパターン A の人の問 2 の回答

ページ 1		ページ 2		ページ 3	
選択した語	スコア	選択した語	スコア	選択した語	スコア
検索エンジン	6	検索エンジン	6	検索エンジン	8
利用浸透	5	Microsoft	5	Microsoft	8
Microsoft	5	Bing	4	Bing	4
Bing	4	検索市場シェア	4	シェア	2
宣伝	2	シェア	4	検索市場	2
宣伝費投入	1				
検索者	1				

表 9 履歴データ 2 におけるパターン B の人の問 2 の回答

ページ 1		ページ 2		ページ 3	
選択した語	スコア	選択した語	スコア	選択した語	スコア
デジタルカメラ機能	7	docomoPROseries	5	NTT ドコモ	12
NTT ドコモ	5	docomoSTYLEseries	5	冬春モデル	6
無線 LAN	4	冬春モデル	5	ラインアップ	6
携帯電話	3	セパレートケータイ	4		
瞬速起動	3	機種	4		
カメラ	2	端末	1		

表 10 履歴データ 3 におけるパターン D の人の問 2 の回答

ページ 1		ページ 2		ページ 3	
選択した語	スコア	選択した語	スコア	選択した語	スコア
スマートフォン	6	iPhone	9	携帯電話	8
携帯電話市場	6	Google ケータイ	7	Android	5
スマートフォン向け OS	4	機能 OS	5	搭載携帯	4
iPhone	4	アップル	3	Nexus One	3
モバイル OS メーカー	3			正式発表	2
iPhone OS	1			グーグル	2

て URL を保存するという方法も今後の課題として考えられる。

今後は、ユーザの Web 閲覧行動を精緻化するとともに、パターンを考慮した推薦法に拡張していく。

## 文 献

- [1] Maria Grineva, Maxim Grinev, and Dmitry Lizorkin. Extracting key terms from noisy and multi-theme documents. *WWW2009*, Vol. Mining for Semantics, pp. 661–670, 2009.
- [2] 陳光敏, 小林亜樹, 山岡克式, 酒井善則. Web コンテンツ間類似度を用いた関連情報探索空間の構成法. 信学技法, No. 2004-02, pp. 19–24, 2004.
- [3] 鶴原翔夢, 高須賀清隆, 丸山一貴, 寺田実. 独立成分分析を用いた Web 閲覧履歴の解析と Web ページ推薦への応用. *DEWS2008*, B2-3, 2008.
- [4] 堀幸雄, 今井慈朗, 中山亮. ユーザの web 閲覧履歴を用いた検索支援システム. 情報知識学会誌, Vol. 17, No. 2, pp. 95–100, 2007.
- [5] 山本覚, 松尾豊. 行動履歴に基づくページ間ネットワークの分析. *JWEIN09*, 2009.
- [6] 岩田具治, 山田武士, 上田修功. 購買順序を効率的に用いた協調フィルタリング. 情報知識学会論文誌, Vol. 49, No. 4, pp. 125–134, 2008.
- [7] 中川裕志, 湯本紘彰, 森辰則. 出現頻度と接続頻度に基づく専門用語抽出. 自然言語, Vol. 10, No. 1, pp. 27–45, 2003.
- [8] 松尾豊, 友部博教, 橋田浩一, 中島秀之, 石塚満. Web 上の情報からの人間関係ネットワークの抽出. 人工知能学会論文誌, Vol. 20, No. 1, pp. 46–56, 2005.

## 6. おわりに

本論文ではユーザが閲覧したページのキーワードを抽出し、ページを移動することによって変化する話題のキーワード遷移からユーザの興味を推定する手法を提案し、さらに提案法を用いた Web 文書推薦システムを実装した。

本システムにおいてキーワード遷移とは各ページに出現するそのページの話題を表すキーワードをノード、キーワード間の *Simpson* 係数の値をエッジの重みとしたグラフによって表わされ、構築されたグラフを解析することによって、ユーザがページを移動するたびにユーザの興味を推定しそれに沿った文書の推薦を行った。

ユーザの Web 探索行動をキーワードのリストの積み重ねで表現することと、*Simpson* 係数によって構築されるキーワード遷移グラフの有効性を検証するために、ユーザの Web 閲覧行動を 4 つのパターンに分けて利用者実験を行った。その結果、キーワードのリストの積み重ねからでもユーザの Web 探索行動の特徴を知ることができ、またシステムがキーワード遷移グラフを用いて提示するキーワードとユーザが選択するキーワードも重なるものがあることが確認できた。