

Web 検索のための質問キーワードの 時間依存性に基づくクラスタリング手法

賀家智代[†] 角谷和俊^{††}

現在、Web 検索のための検索エンジンが普及しているが、検索結果として呈示される Web ページの内容や観点は多岐にわたっていて、重要度の高い Web ページがユーザの意図に合致するとは限らない。そのため Web 検索結果を解析することによって、トピックごとや関連の深いページをクラスタリングして呈示する手法がいくつか提案されている。一般に、同じトピックや関連の深いページであっても質問キーワードに関して異なる観点の Web ページが混在する場合がある。一時点の Web ページのみを解析する手法ではキーワードに関する観点の分析は難しく、このような問題は解決されない。そこで本研究では、時系列的な特性に基づく新たなクラスタリング方式を提案する。Web アーカイブに蓄積された過去の Web ページを利用することによって、質問キーワードに関する観点の違いに基づく URL の分類を行う。手順としては、質問キーワードを含む URL を抽出し、URL 毎にキーワードの出現傾向を解析する。次にキーワードの時間依存性に基づき Web ページを検索し、最後にその結果をクラスタリングして呈示する。本稿では提案する手法について述べ、評価及び分析を検討する。

A Clustering Method Based on the Temporal Relation of Query Keywords for Web Search

TOMOYO KAGE,[†] and KAZUTOSHI SUMIYA^{††}

The Web search engines based on the keywords are popular because it is easy for users to retrieve information. However, retrieved Web pages have various contents and perspectives, and the pages do not match users' intentions because level of importance is high. For the reasons stated above, some clustering methods of topic or relationship are proposed. The clustering methods are divided into methods based on contents and methods based on structure. However, same topic's or related pages may be mixed-up some perspectives about query keywords. It is difficult for the conventional methods analyzing only today to analyze the perspectives about query keywords, and the methods can not solve the problem. In this study, we propose new clustering method based on a temporal aspect. We classify retrieved pages into same perspectives groups about query keywords by Web logs. First, we extract URLs included query keywords and analyze appearance tendencies of the keywords' each URL. Next, we retrieve Web pages based on the temporal relations of query keywords. Retrieved Web pages are clustered based on the temporal relations. In this paper, we describe our proposed method and examine the evaluation and the analysis.

1. はじめに

従来、Web 検索の結果として出力される Web ペー

ジは、キーワードを含み、重要度に基づくランク付けがなされて呈示される。Web ページのランク付けは、大量の検索結果からユーザが効率的に情報取得するために有効な手法であり、様々な方式¹⁾²⁾が提案されている。しかしながら、キーワードを質問とする検索の結果は、単純にキーワードが含まれているだけで内容は多岐にわたっており、ユーザの意図した Web ページが上位にランキングされるとは限らない。つまり、ランクが上位である重要度の高い Web ページでも、ユーザが要求する内容とは一致しない場合がある。このような場合、ユーザは逐次 Web ページを閲覧して検索を行わなければならない、重要度のみでの呈示方式

[†] 兵庫県立大学大学院環境人間学研究所
Graduate School of Human Science and Environment,
University of Hyogo
〒 670-0092 兵庫県姫路市新在家本町 1 丁目 1-12
E-mail: nd05w005@stshse.u-hyogo.ac.jp

^{††} 兵庫県立大学環境人間学部
School of Human Science and Environment,
University of Hyogo
〒 670-0092 兵庫県姫路市新在家本町 1 丁目 1-12
E-mail: sumiya@shse.u-hyogo.ac.jp

では検索支援は難しいと考えられる。

前述した問題の解決方法として、検索結果の Web ページをクラスタリングして呈示する検索エンジンが提供されている^{☆☆☆}。これらの検索エンジンは、キーワードに関する Web ページをトピック毎に呈示することが可能で、クラスタリングすることによってユーザの検索を支援する。しかしながら、同じトピックの Web ページであってもユーザの意図に合致した情報であるとは限らない。例えば、「旅行」という Web ページでも「旅行に行った話を記述している個人のページ」もあれば「旅行情報をまとめたガイドブックのようなページ」、「ツアープランを掲載した旅行会社のページ」など、旅行に関する異なる観点の Web ページが含まれている。つまり、同じトピックでも URL によって観点が異なる。

そこで、我々は質問キーワードに関して Web ページにおけるキーワードの出現履歴を利用して、観点の異なる Web ページを自動的にクラスタリングする手法を提案する。以降、2 節では本研究のアプローチ及び関連研究について述べ、3 節では質問キーワード間の順序関係について説明する。4 節では検索及び結果のクラスタリング手法について論じ、5 節では、プロトタイプシステムと評価実験を検討する。最後に 6 節でまとめと今後の課題を述べる。

2. 本研究のアプローチ

2.1 関連研究

2.1.1 検索結果のクラスタリング

Web 検索結果のクラスタリング手法は、Web ページのテキストを解析する自然言語に基づくアプローチと、リンクの参照関係を解析する Web の構造に基づくアプローチの 2 種類の方法に大別できると考えられる³⁾。

前者としては、成田ら⁴⁾や太田ら⁵⁾によって、内容に基づくクラスタリング方式が提案されている。本研究は現時点の Web ページの内容を分析するのではなく、質問キーワードの過去の Web ページにおける出現傾向を利用する時系列的な観点に基づく手法であるためこれらの研究とはアプローチが異なる。

後者としては、Wang ら⁶⁾や大野ら³⁾によって Web のリンク構造に基づくクラスタリング方式が提案されている。これらはそれぞれ、リンクの共有度に基づく方式、及び最大流アルゴリズムという Web のリンク

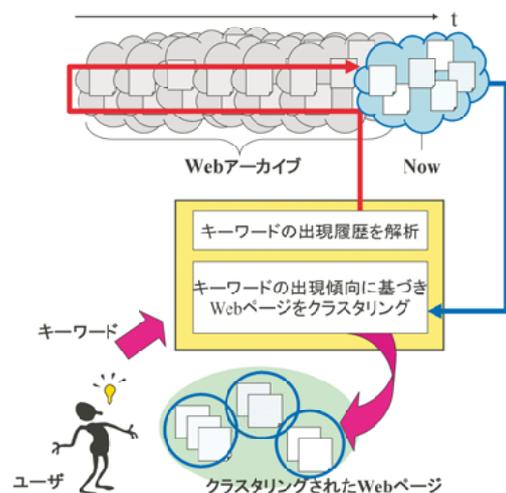


図 1 概念図

構造に着目した方式である。本方式はリンク構造を用いるのではなく、過去の Web ページの傾向から検索結果として得られた URL をクラスタリングするため、これらの研究とは手法が異なる。

2.1.2 履歴を利用した Web 検索方式

履歴を利用した Web 検索方式としては、Web アーカイブを利用した再ランキング方式や検索エンジンに入力される質問キーワードを利用した検索支援方式などが提案されている。

Adam ら⁷⁾⁸⁾によって、Web アーカイブに格納された Web ページの更新履歴を解析し、再ランキングする手法が提案されている。これらの手法は、Web アーカイブを利用して Web 検索結果のランキングを再編するもので、「リンク構造に基づくランキングではユーザの要求している質の良い Web ページが上位にならない場合がある」と明示し、更新状況とその内容の差分を利用することによって質の良いページを上位にランキングする。Web アーカイブを利用して検索結果を変化させるという点で本研究と類似しているが、目的がランキングである点、質問キーワードの意味を一つに捉え、差異を抽出する点で異なる。

Chien ら⁹⁾や Vachos ら¹⁰⁾によって、検索エンジンに入力される質問キーワードの履歴を解析し、関係のある単語を抽出する検索支援方式が提案されている。検索エンジンに入力されるキーワードの普遍的な傾向を解析するこれらの研究に対して、本研究では Web ページの履歴を利用して出現するキーワードの傾向を解析し、URL 毎の傾向を抽出する。

2.2 本研究の概要

我々は、従来の Web クラスタリング手法である「内

[☆] Clusty.com: <http://clusty.com/>

^{☆☆} Clusty.jp: <http://clusty.jp/>

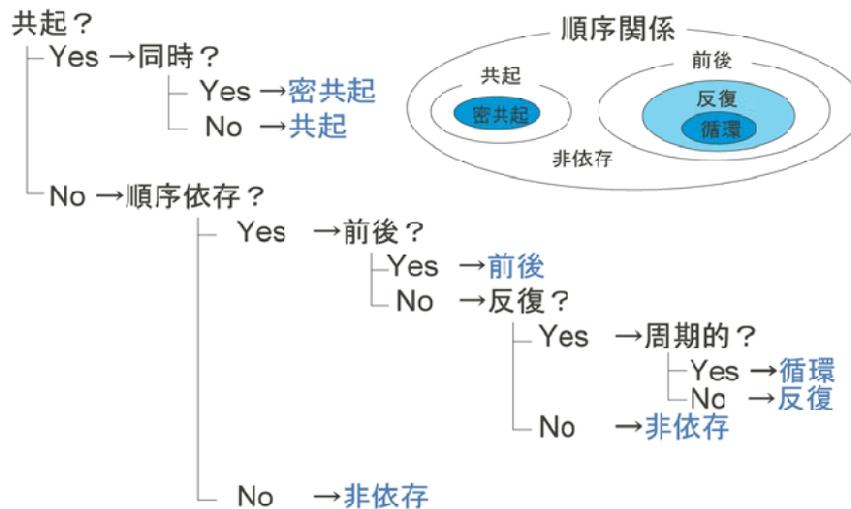


図2 キーワードの順序関係と判定アルゴリズム

容」や「構造」に基づく手法とは異なる時系列的なクラスタリング手法を提案する(図1)。すなわち、本手法ではユーザが入力した質問キーワードとWebページの時間的關係を利用したクラスタリングを行う。特徴としては、Webページにおけるキーワード単独の傾向ではなく、質問キーワードとして入力された複数のキーワード間の関係まで抽出する点にある。我々は、質問キーワードを構成しているキーワード間の関係を判定することによって、Webページの特徴としてより質問キーワードに基づく特徴量を抽出することができる。

以下に本手法の概要を説明する。まず、Webページにおける質問キーワードの出現傾向を解析する。質問キーワードに対する様々な観点のWebページをクラスタリングするために、本方式ではURL毎に解析を行う。次に、キーワードの出現傾向からキーワード間の順序関係を判定する。順序関係とは、キーワードと一緒に出現する「共起」、あるキーワードがあるキーワードの後に出現する「前後」といった関係である¹¹⁾。さらに、順序関係に基づき質問を生成する。ここで生成される質問は、過去のWebページに出現するキーワードの傾向を考慮しているためキーワードを含んでいない有用なWebページを取得するためにOR結合される場合がある。最後に、順序関係に基づき検索されたWebページをクラスタリングして呈示する。この時、クラスタ特有の単語をラベルとして抽出し、クラスタ毎にラベルを付与する。

3. 質問キーワードの順序関係

3.1 キーワードの順序関係

ユーザが検索エンジンに入力した質問キーワードの關係として6つの順序関係を定義する。基本的な順序関係を共起、前後、反復、非依存關係とする。また、基本形の特種な關係として「共起」の強い關係を「密共起」、「反復」の強い關係を「循環」とする。順序關係の概念図を図2の右上部分に示す¹¹⁾。

共起 (co-occurring) 時系列ページにおいてキーワード a, b が共に同時期に出現する關係を共起關係という。一般的な共起は複数のキーワードが1つのWebページに出現することをいうが、本研究では、共起の範囲をページ単位ではなく時間単位とする。

前後 (ordered) キーワード a, b が互いに順序依存している關係、すなわち時系列のWebページにおいて一方が他方より時間的に前に出現する關係を前後關係という。例えば、「牛丼」と「豚丼」が挙げられる。例でも分かるように前後關係は因果關係を包含している。後述する反復については繰り返す前後關係として更に分類される。

反復 (repeated) キーワード a, b が時系列順に反復して出現する關係、すなわち a と b が時系列のWebページに交互に出現する關係を反復關係という。例えば、「応募」と「当選」が挙げられる。後述する循環については厳密な反復關係として更に分類される。

非依存 (independent) 時系列ページにおいてキーワード a, b が独立して出現する關係、すなわちキーワードが互いにランダムに出現する關係を非依存關係

という。例えば、飲食店の URL で「定食」と「限定メニュー」は非依存関係になる。非依存関係は上記で述べた前後、共起、反復関係以外とする。

密共起 (strict co-occurring) 密共起関係は上記で述べた共起関係の中でも更に厳密な関係であり、時系列の Web ページにおいて、キーワード a, b が共に時系列の Web ページの 1 時点のページに出現している関係をいう。

循環 (cyclic) 循環関係は上記で述べた反復関係の中でも更に厳密な関係であり、時系列の Web ページにおいて、キーワード a, b が一定の周期で繰り返して出現する関係をいう。この関係は、反復関係とは異なりキーワードは規則的に出現する。例えば「春」と「秋」は循環関係といえる。

3.2 順序関係の判定

本手法では、ユーザによって入力されたキーワードを 2 つのキーワードの組に分割し、各々の組について順序関係を判定する。手順としては、最初に共起関係か否かを判定する。次に前後関係かどうかを判定し、最後に反復関係かどうかを判定する。判定アルゴリズムを図 2 に示し、判定する任意のキーワードを a, b として、以下に述べる。

3.2.1 共起関係の判定

同時期といえる時区間の閾値を定め、キーワード a, b が閾値時間以内出現している場合を共起とする。時系列の Web ページに出現するキーワードの総数のうち共起している割合が閾値 α 以上である場合を共起関係と判定する。共起関係でない場合は前後関係の判定を行う。共起関係である場合は、更に同時期といえる時区間が 0 の場合の共起の割合を算出し、 α 以上である場合を密共起関係とする。

3.2.2 前後関係の判定

まず、時系列の Web ページにおいて順序依存しているかを判定する。 a から b の順序が成立する区間の総和 ($I_{a \ll b}$)、 b から a の順序が成立する区間の総和 ($I_{b \ll a}$) の 2 種類の区間を抽出し、時系列の Web ページ全区間に占める割合を算出する。その割合が大きい時キーワードは順序依存していると考えられるため、この値が閾値 β より小さければ非依存関係と判定する。大きければ $I_{a \ll b}$ と $I_{b \ll a}$ の時区間の偏りを求め、その偏りの値が閾値 γ 以上であれば前後関係とみなし、閾値 γ 未満であれば反復関係の判定を行う。

3.2.3 反復関係の判定

a から b の順序が成立する区間 ($i_{a \ll b}$) と b から a の順序が成立する区間 ($i_{b \ll a}$) が交互に出現していて、その割合が大きければ反復関係と判定する。つまり、

$i_{a \ll b}$ と $i_{b \ll a}$ が交互に出現している区間を足し合わせ、全区間における割合が閾値 θ より大きければ反復関係とみなす。反復関係でない場合は非依存関係となり、反復関係である場合は次に循環関係であるか否かを判定する。循環関係の場合、 $i_{a \ll b}$ と $i_{b \ll a}$ の各々の時区間が一定であるので、(周期と見なせる) それらの区間の分散値を計算する。 $i_{a \ll b}$ と $i_{a \ll b}$ の分散値が共に閾値 δ よりも小さい場合、循環関係とみなす*。

4. 検索結果のクラスタリング

4.1 順序関係に基づく質問生成

前述した順序関係に基づき、現在の Web ページの検索を行う。検索方式は以下の通りである。

まず、検索の対象となる URL を決定する。本方式は過去の履歴を考慮するため、現時点だけでなく過去に質問キーワードを含んでいる URL も検索対象とする。すなわち、現時点では質問キーワードを含んでいない URL でも、過去の時系列の Web ページ全体に質問キーワードを含んでいるとき、その URL は検索対象となる。

次に、検索対象となった URL からキーワードの順序関係を抽出し、その順序関係に基づき論理式を生成する。 AND や OR に基づく論理式は単純に以下の通りとする。

- 順序依存関係であるキーワードは OR で結合する。
- 順序非依存関係であるキーワードは AND で結合する。

最後に、検索対象である URL に問い合わせる質問を生成する。生成方法は、順序関係に基づく AND や OR で結合された論理式を組み合わせて行う。例えば、{お中元, お歳暮} が反復、{お歳暮, ギフト} が非依存、{ギフト, お中元} が非依存である場合、それぞれの論理式は、(お中元 OR お歳暮)、(お歳暮 AND ギフト)、(ギフト AND お中元) となり、生成される質問は (お中元 AND ギフト)、(お歳暮 AND ギフト) となる。

4.2 クラスタリング手法

順序関係に基づき、検索結果のクラスタリングを行う。本手法では、順序関係の概念構造を利用して階層型のクラスタリングを構築する。例えば、「お中元」と「お歳暮」の 2 つのキーワードを質問キーワードとして検索を行い、その結果をクラスタリングした場合、まず、前後関係または反復関係と判定された「百貨店」の Web ページがクラスタリングされまとめられる。次

* 分散値はキーワードの散ばり具合を表す指標である。

表 1 判定された順序関係とそれに基づくクラスタリング

質問キーワード	URL	順序関係	
{ お中元, お歳暮, ギフト }	http://www.tenmaya.co.jp/ http://www.sanyo-dp.co.jp	反復, 非依存	お中元 ○ お歳暮, { お歳暮, ギフト }, { ギフト, お中元 } お中元 ○ お歳暮, { お歳暮, ギフト }, { ギフト, お中元 }
	http://www.taka.co.jp/ http://www.gift.or.jp/	循環, 反復 反復	お中元 ○ お歳暮, お歳暮 ○ ギフト, ギフト ○ お中元 お中元 ○ お歳暮, お歳暮 ○ ギフト, ギフト ○ お中元
	http://gift.indac.jp/ http://www.suzuto.co.jp/ http://www.oiwai.co.jp/	密共起	お中元 ⊗ お歳暮, お歳暮 ⊗ ギフト, ギフト ⊗ お中元 お中元 ⊗ お歳暮, お歳暮 ⊗ ギフト, ギフト ⊗ お中元 お中元 ⊗ お歳暮, お歳暮 ⊗ ギフト, ギフト ⊗ お中元
	http://www.jtb.co.jp/ http://www.yomiuri-ryokou.co.jp/	反復	桜 ○ 紅葉, 紅葉 ○ 城, 城 ○ 桜 桜 ○ 紅葉, 紅葉 ○ 城, 城 ○ 桜
	http://www.himeji-kanko.jp/ http://www.kobe-photo.com/ http://www.naviu.net/ http://www.city.tatsuno.hyogo.jp/	密共起 共起	桜 ⊗ 紅葉, 城 ⊗ 紅葉, 城 ⊗ 桜 桜 ⊗ 紅葉, 城 ⊗ 紅葉, 城 ⊗ 桜 桜 ⊗ 紅葉, 城 ⊗ 紅葉, 城 ⊗ 桜 桜 ⊕ 紅葉, 城 ⊕ 紅葉, 城 ⊕ 桜

に、それらの Web ページは、最近お中元やお歳暮を提供し始めた「前後関係」であるページとお中元やお歳暮を常に提供している「反復関係」のページに分類される。このようにクラスタリングを行うことによって、大手と新しい店の Web ページを分類することができる。つまり、クラスタリング手法としては、まず最初に大きな大別として「共起または密共起」「前後または反復または循環」「非依存」に分類する。次に、「共起または密共起」のクラスタを「共起」と「密共起」に、「前後または反復または循環」を「前後」と「反復または循環」に分類する。最後に「反復または循環」を「反復」と「循環」に分類する。このようにして、階層を持たせたクラスタリングを行う。

クラスタを生成した後、どのクラスタの Web ページを閲覧するかユーザが判断しやすいように、クラスタ毎にラベルを付与する。単語抽出の対象となる Web ページは検索結果として呈示される Web ページとする。ラベルはクラスタを特徴付ける単語で、単純に $tf-idf$ 値の上位とする。すなわち、あるクラスタ c に含まれる単語 w の数を n_{cw} 、 c を含む検索結果として生成された N 個の全てクラスタ C のうち w を含むクラスタ数を N_w とすると、あるクラスタ c に含まれる単語 w の $tf-idf$ 値は、

$$n_{cw} \log \frac{N}{N_w} + 1 \quad (1)$$

となる。これらは階層毎に行うものとする。

例えば、{お中元, お歳暮, ギフト} が質問キーワードである場合、{お中元, お歳暮} が循環関係、{お歳暮, ギフト} と {ギフト, お中元} が非依存 (つまり「ギフト」が非依存) であるクラスタに含まれる Web ページは、特有の時期 (おそらく夏) に「お中元」や (おそらく冬に)「お歳暮」が出現していると考えられ、百貨店などのページであると考えられる。そのため、循環関係のクラスタとしては「店」などの単語が抽出

され、ラベルが付与される。また、{お中元, お歳暮} が循環、{お歳暮, ギフト} と {ギフト, お中元} が反復であるようなクラスタに含まれる Web ページは、おおよそギフトの情報を提供してお歳暮とお中元がタイムリーに出現しているようなギフト専門のページであると考えられる。そのため、シーズンに依存した行事に用いられるような「包装」や「ラッピング」などの単語が抽出され、ラベルが付与される。さらに、{お中元, お歳暮}, {お歳暮, ギフト}, {ギフト, お中元} の全ての組が密共起であるようなクラスタに含まれる Web ページは、常にギフトの情報を提供しているようなページで、常にギフトを贈るという観点での「カタログ」などの単語が抽出され、ラベルが付与される (表 1)。

5. プロトタイプシステムと実験

5.1 プロトタイプシステム

システム構成を図 3 に示す。システムを構成するユニットについて以下に説明する。

- (1) **Web ページ収集部** ユーザによって入力されたキーワードを含む Web ページの URL を取得し、InternetArchive(http://web.archive.org/web/*/(任意の URL)) から時系列の Web ページを自動収集する。ただし、URL が変更されることを考慮して 2 リンク先の同一サイト内の Web ページを同一 URL と見なして収集する。
- (2) **インデックス生成部** Web ページが収集された時間、その Web ページに含まれる単語、ある URL の時系列 Web ページ全体に含まれる単語をインデックスとする。単語は、Web ページに含まれるテキストを形態素解析して抽出する。ページが収集された時間は InternetArchive(<http://web.archive.org/web/>(14桁の数値)/任意の URL)) の数値部分から取得

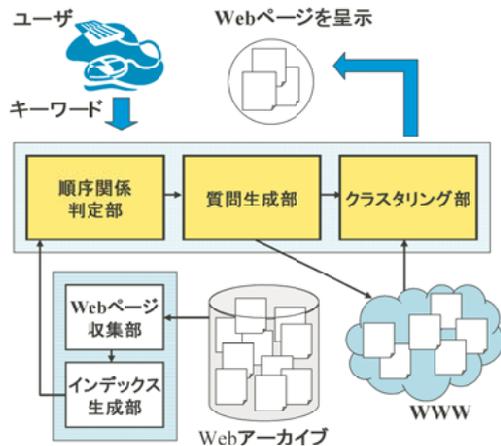


図3 システム構成図

する*。

- (3) **順序関係判定部** 質問キーワードを2つずつに分解し、各々の組に対して順序関係を判定する。
- (4) **質問生成部** 順序関係を基に質問生成し、その質問により問い合わせを行う。
- (5) **クラスタリング部** 順序関係に基づき、検索結果をクラスタリングしてユーザに提示する。出力イメージとしては、ユーザが順序関係を指定するインタラクティブ提示(図4)や自動的にクラスタリングして提示するデフォルト提示(図5)などのインタフェースが考えられる。

5.2 実験

順序関係に基づく Web ページのクラスタリングを評価するために実験を行った。なお、閾値は $\alpha, \beta, \gamma, \theta$ をそれぞれ 0.3, 0.6, 5, 0.9, 共起の範囲を 90 日とした。結果として、観点の異なる Web ページにクラスタリングすることができた(表1)。なお、表1ではキーワード a と b の順序関係を記号で表している。順序関係とその記号は以下のとおりである。

共起: $a \oplus b$, 密共起: $a \otimes b$

前後: $a < b$, 反復: $a \circ b$, 循環: $a \odot b$

非依存: $\{a, b\}$

[お中元, お歳暮, ギフト] まず, {お中元, お歳暮, ギフト} を質問キーワードとしてクラスタリングを行った。クラスタは3つ生成された。Web ページを分析した結果と考察を述べる。

- 1つめのクラスタ(表1上2つのURL)に含まれる Web ページは、デパートのページとなった。

* サイト内のページに付与される時間データはトップページに依存する。



図4 出力画面イメージ(インタラクティブ提示)



図5 出力画面イメージ(デフォルト提示)

周期的に「お中元」と「お歳暮」が出現すると考えられるため循環関係になると思われたが、Web アーカイブの取得状況に依存して反復関係となった。しかしながら、同じ「店」という単語が特徴的な Web ページがまとめられた。

- 2つめのクラスタ(表1の上から3つ目と4つ目のURL)に含まれる Web ページは、ラッピング等の紙を製造している企業のページとギフト専門のページとなった。どちらも特徴的に「ラッピング」という単語が出現していることから、ラッピングに関して有用な情報を含むページとして共通点があった。また、このクラスタは階層構造をしていた。同じ「ラッピング」という情報を含む Web ページであるが、「包み紙」専門の URL と「ギフト」専門の URL に分けられた。
- 3つめのクラスタ(表1の上から5~7番目のURL)に含まれる Web ページは、ギフトショップのページとなった。「カタログ」という単語が特徴的に出現していたため、ギフトカタログに関し



図 6 反復関係と判定された Web ページ (2004.11)



図 7 密共起と判定された Web ページ (2004.11)

て有用な情報を含むページとして共通点があった。[桜, 紅葉, 城] 次に, {桜, 紅葉, 城} を質問キーワードとするとクラスタは2つ生成された。なお, 2004年11月を現在としてクラスタの異なる Web ページ2枚を図6^{*}, 7^{**}に示す。結果と考察を以下に述べる。

- 1つめのクラスタ (表1下から6つ目と5つ目のURL) に含まれる Web ページは旅行会社のページとなった。周期的に「桜」と「紅葉」が出現すると思われたが, 反復関係となった。しかしながら, 同じ「旅行」や「ツアー」といった共通の Web ページがまとめられた。ただし, これらの Web ページは現時点が春または秋でなければ呈示されることはない (図6)。
- 2つめのクラスタ (表1の下4つのURL) に含まれる Web ページは, ローカルな情報を提供しているページとなった。これらの Web ページはその地域の年中行事や様子をまとめているような Web ページがい (図7)。

6. おわりに

我々は, 過去の Web ページにおける質問キーワードの出現履歴に基づく Web 検索のためのクラスタリング手法を提案した。質問キーワードの出現履歴を解析して, キーワード間の順序関係を判定し, 順序関係の階層を利用して階層的な Web ページのクラスタリングを行った。本手法は, 従来のトピックを分類するだけのクラスタリング方法とは異なり, 同じトピックでも観点の異なる Web ページに分類することができ, 実験で有用性が認められた。

^{*} <http://www.jtb.co.jp/>

^{**} <http://www.kobe-photo.com/>

今後の課題としては, より複雑な順序関係の判定方式の確立と更に順序関係に準ずる質問生成方式の提案などが挙げられる。

謝 辞

本研究の一部は, 平成18年度科研費基盤研究(B)(2)「Web アーカイブと映像アーカイブを融合した次世代デジタル・ライブラリに関する研究」(課題番号: 16300028)によるものです。ここに記して謝意を表すものとします。

参 考 文 献

- 1) Brin, S. and Page, L.: The anatomy of a large-scale hypertextual Web search engine, *Proceedings of the 7th International World Wide Web Conference(WWW1998)*, Brisbane, Australia, pp. 107-117 (1998).
- 2) Kleinberg, J.: Authoritative sources in a hyperlinked environment, *Journal of ACM*, Vol. 48, pp. 604-632 (1999).
- 3) 大野成義, 渡辺匡, 片山薫, 石川博, 太田学: Max Flow アルゴリズムによる Web ページのクラスタリング方法, 夏のデータベースワークショップ DBWS'05, 電子情報通信学会, 情報処理学会 (2005).
- 4) 成田宏和, 太田学, 片山薫, 石川博: Web 文書検索のための非排他的クラスタリング手法の提案~NOCTURNE(New Overlapping Clustering Tool Using Ranking iNformation of search Engine), 第14回データ工学ワークショップ DEWS'03, 電子情報通信学会 (2003).
- 5) Ohta, M., Narita, H. and Ohno, S.: Overlapping Clustering Method Using Local and Global Importance of Feature Terms at NTCIR-4 Web Task, *Working Notes of the*

- 4th NTCIR Meeting, Supplement volume 1*, pp. 102–110 (2004).
- 6) Wang, Y. and Kitsuregawa, M.: Use Link-based Clustering to Improve Search Results, *Proceedings of the 2nd International Conference on Web Information System Engineering*, IEEE Computer Society (2001).
 - 7) Jatowt, A., Kawai, Y. and Tanaka, K.: Temporal Ranking of Search Engine Results, *Proceedings of the The Fifth International Conference on Web Information Systems Engineering (WISE2005)*, pp. 43 – 52 (2005).
 - 8) Jatowt, A., Kawai, Y. and Tanaka, K.: Using Web Archive for Improving Search Engine Results, *Proceedings of The Eighth Asia Pacific Web Conference (APWeb2006)*, pp. 893 – 898 (2006).
 - 9) Chien, S. and Immorlica, N.: Semantic Similarity Between Search Engine Queries Using Temporal Correlation, *Proceedings of the 14th International Conference on World Wide Web (WWW2005)*, pp. 2 – 11 (2005).
 - 10) Vlachos, M., Meek, C., Vagena, Z. and Gunopulos, D.: Identifying Similarities, Periodicities and Bursts for Online Search Queries, *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD2004)*, pp. 131 – 142 (2004).
 - 11) 賀家智代, 角谷和俊: 質問キーワードの順序依存性に基づく Web アーカイブ検索方式, 第 17 回データ工学ワークショップ DEWS'06, 電子情報通信学会 (2006).