Webアーカイブを用いた時系列パターンに基づく検索支援方式

小野田　透†　　賀家　智代††　　角谷　和俊†††

† 姫路工業大学環境人間学部　〒 670-0092 兵庫県姫路市新在家本町 1 丁目 1-12
†† 兵庫県立大学大学院環境人間学研究科　〒 670-0092 兵庫県姫路市新在家本町 1 丁目 1-12
††† 兵庫県立大学環境人間学部　〒 670-0092 兵庫県姫路市新在家本町 1 丁目 1-12
E-mail: †{na03q040,nd05w005}@stshse.u-hyogo.ac.jp, ††sumiya@shse.u-hyogo.ac.jp

あらまし　近年，多くの人々が日々の情報収集の手段として Web を活用するようになってきている．ユーザは検索エンジンを用いて質問キーワードを入力することで様々な情報を得ることが出来るようになった. しかしながら，現在のキーワード検索ではユーザが複数の質問キーワードを入力し検索を行ったとき，質問キーワードが全て出現しているページでなければ検索結果として取得されにくい．本研究では，Web アーカイブを用いて Web ページの過去における質問キーワードの出現状況を分析することで，現在のページで質問キーワードが全て出現していないページであっても，ユーザにとって有用と思われるページであれば取得を行う．さらに，質問キーワードの時系列的な出現パターンを利用して適切な過去のページを抽出し，現在のページと共にユーザに提示することでよりユーザの意図に合った結果を提示する方式を提案する．

キーワード　Web とインターネット，情報検索，時空間 DB，Web アーカイブ

A Retrieval Method based on Temporal Pattern using Web Archives

Toru ONODA†, Toyomo KAGE††, and Kazutoshi SUMIYA†††

† School of Humanities for Environmental Policy and Tecnology, Himeji Institute of Techology
1-1-12 Shinzaike-honcho, Himeji, Hyogo 670-0092, Japan
†† Graduate School of Human Science and Environment, University of Hyogo
1-1-12 Shinzaike-honcho, Himeji, Hyogo 670-0092, Japan
††† School of Human Science and Environment, University of Hyogo
1-1-12 Shinzaike-honcho, Himeji, Hyogo 670-0092, Japan
E-mail: †{na03q040,nd05w005}@stshse.u-hyogo.ac.jp, ††sumiya@shse.u-hyogo.ac.jp

Abstract　Recently, a lot of people become use the internet as a means of the information gathering. The user came to be able to obtain various information by inputting the question keyword by using the search engine. However, it is difficult to get Web pages that all keywords do not appear when the user inputs two or more keywords. In this study, We are analyzed the temporal pattern of keywords on the Web pages in the past with Web archive. As a result, if it is a useful page for the user, it acquires it, even if it is a page where all keywords do not appear on a present page. In addition, It presents a past appropriate page is extracted with a present page to using temporal pattern of the keywords. As a result, it proposes the method to present the result of suitable for the user's intention.

Key words　Web and Internet, Information retrieval, Spacio Temporal DB,Web Archive

1. はじめに

Web を用いた情報収集が一般的なものとなり，中でも Google(注1) や Yahoo!(注2) などで提供されているキーワード検索機能が多用されている．現在のキーワード検索では，ユーザが複数の質問キーワードを入力し検索を行った際，ユーザの入力した質問キーワードが全て出現しているページを検索結果として取得する．例えば，ユーザが複数の質問キーワード「桜」と「紅葉」を入力した場合,「桜」と「紅葉」が共に出現しているページを検索結果として取得する．この場合，キーワード「桜」のみ，あるいはキーワード「紅葉」のみが出現しているページ

(注1)：http://www.google.co.jp/
(注2)：http://www.yahoo.co.jp/

は取得されにくい．

しかし，Web ページの中には春には「桜」の情報を掲載し，秋には「紅葉」の情報を掲載するというように，キーワードが同時に出現せず，時系列的に出現するページが存在する．このようなページはユーザにとって有用である可能性があるが，現在のキーワード検索では取得されにくい．こうしたページは，現在のページと過去のページに質問キーワードが時系列的に出現しているページとして捉えることが出来る．

本方式では，このように質問キーワードが時系列的に出現しているページのなかから，ユーザにとって有用と思われるページを取得する．さらに，取得したページの中から適切な過去のページを抽出し，現在のページと共にユーザに対して提示することで，現在のページを補完する方式を提案する．本方式によってキーワード「桜」とキーワード「紅葉」が時系列的に出現する Web ページを取得することが可能となり，過去の内容を用いて現在の内容を補完することで，よりユーザの意図に合致するページを提示することが可能となる．

本稿では，第 2 節で本研究の概要と関連研究について述べ，第 3 節でキーワードが時系列的に出現するページを抽出する手法について述べる．第 4 節ではキーワードの時系列的な出現パターンの定義とその判定方法について述べる．第 5 節では定義したキーワードの時系列的な出現パターンに基づき，補完ページとして適切な過去のページを抽出する方式について述べる．第 6 節では提案方式のプロトタイプシステムの設計と実験について述べ，第 7 節でまとめと今後の課題について述べる．

2. 本研究のアプローチ

2.1 本研究の概要

本方式では，これまでのキーワード検索では取得されにくかった質問キーワードが時系列的に出現している Web ページを取得する．さらに，質問キーワードの時系列パターンに基づいて適切な過去のページを抽出し，現在のページと共にユーザに提示することで補完を行う．ページの補完によって，よりユーザの意図に合致した内容を提示することが可能となる．

本方式の概要を図 1 に示す．まず，ページの取得は従来のキーワード検索を用いて行い，質問キーワードが時系列的に出現するページを取得するため，検索条件の緩和を行う．ユーザの入力した質問キーワードが a, b のとき，従来のキーワード検索では $a \wedge b$ という質問を生成し検索を行う．その場合，現在のページ内容にキーワード a が出現し，過去のページ内容にキーワード b が出現しているようなページは取得されにくい．よって本方式では，$a, b, a \wedge b$ の 3 つの質問を生成して検索を行い，その結果を取得する．

次に，検索結果として得られた Web ページについて，それぞれの過去のページを Web アーカイブを用いて取得する．過去のページの取得によって，同一の Web ページについて異なる時間のページを得ることが出来る．本稿では，このような同一の Web ページの時系列的なページ集合を「時系列ページ」と呼ぶ．全ての時系列ページにおいて質問キーワードの出現傾向を分析し，質問キーワードが全て出現しているものだけを抽出する．

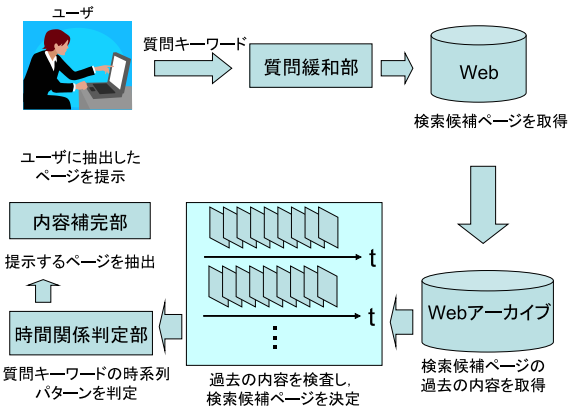

図 1　概　要　図

こうして取得した時系列ページは時系列的に質問キーワードを含んでいるページである．しかし，この時系列ページの現在のページをそのままユーザに提示しても，現在のページにおいて質問キーワードが全て出現していない場合があり十分とはいえない．よって，過去のページを用いて現在のページを補完することでユーザにとってより有用な情報として提示する．このとき，現在のページに対する補完ページとして有効な過去のページを抽出する手法が必要となる．

そこで，本研究では時系列ページにおける質問キーワードの出現パターンを解析し，そのパターンによって補完ページとして有効な過去のページを抽出する．キーワードの時系列的な出現パターンとは，例えば「キーワード a, b は周期性を持って交互に出現する」というパターンである．出現パターンは前後，共起的反復，排他的反復，時間的非依存の 4 つを定義し，時間的非依存はキーワードの出現に意味のあるパターンが無いものと定義する．よって，前後，共起的反復，排他的反復のいずれかの関係と判定された時系列ページに対して各パターンの特性に基づいて過去のページの抽出を行い，補完ページとして現在のページと共に提示する．

2.2 関 連 研 究

Adam [1] [2] らは，Web アーカイブを用いて Web ページの更新履歴を解析し，ページを再ランキングする方式を提案している．これらの手法は，Web アーカイブを利用して Web 検索結果の再ランキングを行うもので，過去の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．本研究とは Web アーカイブを用いて過去の Web ページを利用する点で類似しているが，目的がランキングである点，更新状況とその内容の差分を利用する点で異なる．

キーワードの時系列的なパターンを扱うものとして，賀家ら [3] [4] の研究がある．これらは Web アーカイブの検索や検索結果のクラスタリングのために時系列パターンの抽出を行っており，補完ページの抽出を目的としている本研究とは異なる．

湯本ら [5] [6] は Web ページに対して情報の補完を行う研究として，検索結果として提示された複数の Web ページを組み合わせてユーザに提示する方式を提案している．池田ら [7] はユーザの閲覧している Web ページについて，リンク先のペー

ジ情報と内容の類似したページの情報を解析することで，現在ユーザが閲覧しているページの周辺空間を認知しナビゲーション情報を提示する方式を提案している．また，郡ら [8] は Web 閲覧中のユーザの興味のある内容を補完する情報を Blog から抽出して提示することにより，閲覧中のコンテンツの位置づけ，評判などの情報を取得しつつ Web を閲覧する方式を提案している．これらは異なる Web ページの情報を用いて補完を行うという点で本研究と類似しているが，その対象は現在の Web に限られており，過去のページを用いて補完を行う本研究とは異なっている．

3. 検索候補の抽出

3.1 検索条件の緩和

質問キーワードを時系列的に全て含む可能性のある Web ページを取得するため，検索条件を緩和する．ユーザがキーワード a, b, c を入力した場合，従来の一般的なキーワード検索では $a \wedge b \wedge c$ という質問を生成して検索を行う．しかし，それだけではキーワードが時系列的に出現しているページを取得することは難しい．なぜなら，時系列的なページでは，必ずしも現在の内容にキーワード a, b, c が全て出現している必要は無いからである．よって，本方式ではキーワードを時系列的に含むページに対し，それらを取得するための質問を新たに生成する．

まず入力されたキーワードを一語ずつに分解し，それらの全ての組み合わせを生成する．キーワードが複数となる組み合わせでは，キーワード同士を AND で結合し検索質問とする．入力キーワードが a, b, c の場合に生成される質問は，$a, b, c, a \wedge b, b \wedge c, a \wedge c, a \wedge b \wedge c$ の 7 通りである．

3.2 検索候補の決定

生成した質問によって取得した Web ページ集合から，質問キーワードが全て出現している時系列ページを抽出する手法について述べる．まず，一つ一つのページの過去のページを Web アーカイブを用いて取得し，時系列ページの集合を得る．全ての時系列ページについて質問キーワードが出現しているかどうかを調べ，全ての質問キーワードが出現している時系列ページを検索候補として抽出する．

質問キーワードが a, b, c であるときの例を図 2 に示す．図中の長方形が Web ページを表し，その中の文字はページに出現しているキーワードを表している．また，横線で結ばれているページは同一 URL の Web ページである．ページは図の左に配置されているページほど時間的に古いページとなっている．まず，現在のページにおいて，キーワード a, b, c が出現しているかを調べる．ここで，現在のページにキーワード a が含まれているページの場合，過去のページにおいて少なくともキーワード b, c が出現していなければならない．全ての質問キーワードが出現している時系列ページを検索候補として抽出し，キーワードが全て出現していない時系列ページはノイズとして抽出は行わない．同様の手順で全ての時系列ページについて調べ，検索候補集合を得る．

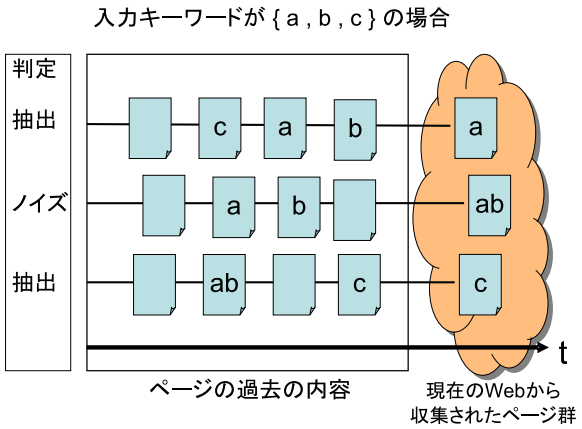

図 2 検索候補抽出

4. 時系列パターンの定義と判定

4.1 時系列パターンの定義

本方式では，過去のページから現在のページの補完に有効なページを抽出するために，キーワードの出現パターンを解析する．こうしたキーワードの時系列的な出現パターンを本稿では時系列パターンと呼ぶ．時系列パターンは無数に存在するが，本方式では以下のパターンを定義する．各パターンの概念図を図 3 に示す．

(a) 共起的反復 ユーザの入力した質問キーワードを a, b としたとき，キーワード a, b が出現しているページが常に時間的に近傍に出現し，それがある程度の周期性を持って繰り返される出現パターンを共起的反復とする．キーワードの共起とは，一般的にはキーワードが同一のページ中に出現する場合を指すが，ここでは共起の範囲を時間的に拡張し，キーワードが同一ページ内に存在していなくても，時間的に近傍に出現していれば共起していると見なす．例えば観光情報について紹介している Web ページにおいて,「桜」と「花見」というキーワードが毎年 4 月に出現するというような場合，共起的反復である．

(b) 前後 ユーザの入力した質問キーワードを a, b としたとき，時系列ページ中のキーワード a, b が出現している時区間において，a, b の順序関係が一意に決定するようなパターンを前後とする．キーワード a, b が前後と判定され，a が b よりも時間的に前のページに出現している時，b が出現したページの後に a が出現することはない．例えば，あるページが小泉が総理在任中は小泉についての情報を掲載し，安倍が総理に就任してからは掲載する情報を安倍に関する情報に切り替えた場合,「小泉」と「安倍」は前後となる．

(c) 排他的反復 ユーザの入力した質問キーワードを a, b としたとき，a, b が交互に出現し，それがある程度の周期性を持って繰り返される出現パターンを排他的反復とする．例えば旅行情報のページにおいて，毎年 4 月に「桜」，10 月に「紅葉」というキーワードが出現するというような場合，排他的反復であるといえる．

(d) 時間的非依存 上記の関係のいずれにも当てはまらない関係を時間的非依存とする．時間的非依存となる時系列ページは，

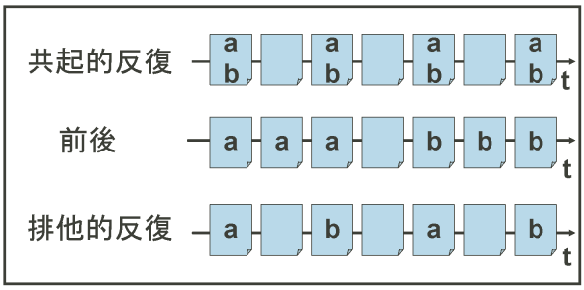

図 3　時系列パターン

キーワードは時系列的に出現しているが，出現パターンの傾向を見つけることが難しいページである．

4.2　時系列パターンの判定

時系列パターンの判定は，ユーザによって入力された質問キーワードの中から 2 つを選択し，その全ての組み合わせについて行う．つまり，質問キーワードが $\{a, b, c\}$ であった場合，$\{a, b\}$，$\{b, c\}$，$\{a, c\}$ の時系列パターンについて判定を行う．判定はそれぞれの質問キーワードが全て出現している時系列ページに対して行う．質問キーワードの時系列パターンを判定するため，前処理として判定を行う時系列ページにおいて質問キーワードが出現している時区間を抽出しておくこととする．

4.2.1　共起的反復の判定

キーワードの時系列パターンが共起的反復であるページを判定する方式について述べる．まず，キーワード a, b が出現している区間から，共起と判定される範囲に存在する a, b を抽出する．その際，a で始まり b で終わる区間 $i_{a \prec b}$ と，b で始まり a で終わる区間 $i_{b \prec a}$ を抽出する．次に，抽出した区間で時系列的に隣り合う $i_{a \prec b}$ と $i_{b \prec a}$ をグルーピングして共起区間を抽出する．

一方，共起区間にはならなかった区間で時系列的に隣り合う $i_{a \prec b}$ と $i_{b \prec a}$ をグルーピングし，共起ではない間隔で a と b が出現している区間も抽出する．そして，共起区間のグルーピング回数と非共起区間のグルーピング回数を算出する．さらに，

- 共起区間生成時のグルーピング回数と非共起区間生成時のグルーピング回数の比率が閾値 α 以上
- 共起区間生成時のグルーピング回数が多く，かつ共起区間と非共起区間が繰り返されている

以上の条件を満たすならば共起的反復の判定を行う．判定は共起区間，非共起区間の各々の分散値を用いて行う．共起区間，非共起区間の分散値が 10^p 以下である場合は共起的反復と判定され，10^p 以上である場合は時間的非依存と判定される．

4.2.2　前後の判定

最初に，判定を行う時系列ページにおいてキーワードに時系列的な順序関係があるかを判定する．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をもって出現していると考えられるため，この値が閾値 β より小さければ時間的非依存と判定する．値が閾値より大きければ $I_{a \ll b}$ 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判定し，閾値 γ 未満であれば排他的反復の判定を行う．

4.2.3　排他的反復の判定

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 $(i_{b \ll a})$ が重複せずに，周期的に出現している場合，排他的反復の判定を行う．まず，ノイズを除去するために前処理を行う．$I_{a \ll b}$ と $I_{b \ll a}$ の重複区間が閾値 δ 以上である場合と，$i_{a \ll b}$ と $i_{a \ll b}$ の分散値が 10^p オーダ以上の場合のどちらか一方にでも該当する場合は時間的非依存となる．

どちらの条件にも該当しない場合は排他的反復の判定を続行する．手順としては，$i_{a \ll b}$ と $i_{b \ll a}$ が交互に出現しているかどうかを判定する．$i_{a \ll b}$，$i_{b \ll a}$ 全てが交互に出現していれば排他的反復とみなす．

5.　時系列パターンに基づく検索支援

前節で，検索候補として出力された時系列ページに対しキーワードの出現に時系列的なパターンがあるか否かの判定を行う方法について述べた．判定の結果，キーワードの時系列パターンが共起的反復，前後，排他的反復のいずれかであった場合，判定された時系列パターンに基づいて過去のページを抽出し，現在のページと共にユーザに提示する．

本節では，過去のページから適切な補完ページを抽出する手法について述べる．キーワードが周期的に出現する排他的反復，共起的反復では，時系列ページ中に存在するキーワード出現区間の中から，補完ページを抽出するのに適した区間を決定する．区間の長さの平均値を基準値とし，基準値に近い値を持つ区間ほどページの抽出に適した区間としてスコアが高くなるようスコアリングを行う．

全ての区間のスコアを算出した後，各キーワード出現区間中のページ一つ一つに対してスコアリングを行う．ページの持つスコアと，ページを含んでいる区間のスコアを積算したものが最終的なページのスコアとなり，スコアの高いページほど補完に適したページとする．

5.1　時系列パターンに基づく時区間抽出

各時系列パターンにおいて補完に適切なページを抽出するため，予備実験として時系列ページの分析を行った．その結果，キーワードが共起的反復，排他的反復のパターンを持つ時系列ページでは，1 つのキーワード出現区間においてキーワードが出現してから時間が経過しているページの方が，キーワードが出現して間もないページに比べ，キーワードに関係する内容が充実している傾向があった．このような傾向は時期に依存したイベントなどの情報を掲載するページで多くみられた．

「スキー」というキーワードに注目したとき，毎年冬にスキーツアーの紹介を掲載する Web ページでは，スキーのシーズンが本格化する前にツアーの予約を始めることから，キーワードの出現時にはあまりページ中にスキーに関する内容は多くない．そして，キーワードの出現からある程度時間が経過し，スキーのシーズンが本格化するとページ中のスキーに関する内容も増加する．このようなページから「スキー」に関する内容を補完するページを抽出するとき，キーワード出現直後の時点の

ページよりも出現から時間が経過している時点のページを補完するべきであると考えられる．よって，本手法では共起的反復，排他的反復の時系列パターンを持つページでは，キーワード出現時から時間が経過しているページほど，つまり一つのキーワード出現区間中で時間的に後ろに位置しているページほど重要性が高いと仮定する．

5.1.1 共起的反復における補完ページ抽出

- 区間のスコアリング

まず，キーワード *a,b* が出現している区間を抽出する．キーワード a, b の共起区間を i_n とし，時間的に前に存在する区間から i_1，i_2 $\cdots$ とする．全ての区間を抽出した後，i_n の長さ $d(i_n)$ を算出する．各ページは Web アーカイブに収集された年月日を時間情報として所持しており，区間の始端にあたるページ P_{start}, 終端にあたるページ P_{end} のもつ時間情報から区間の長さを算出する．長さの単位は日とする．$d(i_n)$ は次の式で求められる．t( ) はページの時間情報を返す関数である．

$$d(i_n) = t(P_{end}) - t(P_{start}) \tag{1}$$

全ての区間における $d(i_n)$ の合計値と総区間数 n の商を計算することで $d(i_n)$ の平均値 θ を求めることが出来る．この θ の値を $d(i_n)$ の基準値とする.θ は次の式で求められる．

$$\theta = \frac{\sum_{k=1}^{n} d(i_k)}{n} \tag{2}$$

θ と $d(i^n)$ の差の絶対値を求め，その値と θ の逆数を積算することで値を正規化し，各区間のスコアを算出する．各区間のスコアは次の式で求められる．

$$score(i_n) = 1 - \frac{\theta - d(i_n)}{\theta} \tag{3}$$

スコアが大きい区間ほど補完ページの抽出に適した区間とする．なお，この式では θ の値に対し $d(i_n)$ があまりに大きな値をとる場合，区間のスコアがマイナス値となる．その場合，スコアがマイナス値となった区間をノイズと見なし，ページの抽出対象から除外する．

- ページのスコアリング

各区間中のページに対してスコアリングを行う．抽出されたキーワード共起区間中に存在するページにおいて，ページごとのスコアを算出する．

まず，各キーワード出現区間中のページの時間的な位置によって各ページのスコアリングを行う．本方式では，キーワードの出現から時間が経過しているページ，つまりキーワード出現区間において時間的に後ろに存在するページほどスコアを高く設定する．各ページのスコアは次の式で求められる．

$$score(P_m) = \frac{t(P_m) - t(P_{start})}{t(P_{end}) - t(Pstart)} \tag{4}$$

スコアの算出はキーワード a, b それぞれに行い，キーワード a, b を共に含むページではキーワード a の出現区間におけるスコア (以降，$score(a)$ とする) にキーワード b の出現区間におけるスコア (以降,$score(b)$ とする) を加算した値をページの持つスコアとする．これは，共起的反復においてはキーワードに関連性がある場合が多く，キーワードが共に出現しているページが重要という考えによる．

次に，補完の対象である現在のページの内容を考慮したスコアの補正を行う．仮に現在のページではキーワード a のみが出現していた場合，重要度が高いのはキーワード b に関する内容を持ったページであるといえる．よって，$score(b)$ に 2.0 を積算する．

現在のページでキーワード a, b が共に出現していた場合，あるいは共に出現していなかった場合は補正は行わない．最後に，各ページが持つ $score(a)$, $score(b)$ の合計値に，ページを含んでいる区間のスコアを積算することで最終的な各ページのスコアを算出し，補完ページとして全キーワード共起区間に含まれるページ数の 30% をスコアが高いものから順に抽出する（小数点以下は切り捨て）.

5.1.2 排他的反復における補完ページ抽出

- 区間のスコアリング

まず，キーワード *a,b* が出現している区間を抽出する．次に，補完の対象である現在のページにおけるキーワードの出現状況によって補完ページを抽出する区間を決定する．仮に現在のページではキーワード b が出現していた場合は，キーワード a の出現区間を補完ページ抽出の対象とし，キーワード b の出現区間は対象としない．

キーワード a の出現区間を i_n とし，時間的に前に存在する区間から i_1，i_2 $\cdots$ とする．次に，i_n の長さ $d(i_n)$ を算出する．$d(i_n)$ は式 (1) により求められる．$d(i_n)$ の合計値とキーワード a の総出現区間数 m の商を計算することで $d(i_n)$ の平均値 θ を求め,$d(i_n)$ の基準値とする．θ は式 (2) により求められる．

さらに，θ と $d(i^m)$ の差の絶対値を求め，その値と θ の逆数を積算することで値を正規化し，各区間のスコアを算出する．各区間のスコアは式 (3) により求められる．なお，スコアがマイナス値となった場合，その区間をノイズと見なしページの抽出対象から除外する．

- ページのスコアリング

各区間中のページに対してスコアリングを行う．各キーワード出現区間中の時間的な位置によってページ毎のスコアリングを行い，キーワード出現区間の後ろに存在するページほどスコアを高く設定する．ページのスコアは式 (4) により求められる．

各ページの持つスコアと，区間の持つスコアを積算した値を最終的なページのスコアとし，補完ページとして抽出対象となったキーワードの全出現区間に含まれるページ数の 30% をスコアが高いものから順に抽出する（小数点以下は切り捨て）.

5.1.3 前後における補完ページ抽出

前後は上記の二つの時系列パターンとは性質の異なる特殊な時系列パターンで，キーワード *a,b* は因果関係を持つと仮定する．キーワード a, b が前後関係であり，a が b よりも時間的に前に出現している場合，まず a, b の出現区間をそれぞれ抽出する．このとき，例外的なパターンとしてキーワード a の出現区間とキーワード b の出現区間に重複が生じる場合，重複区間に存在するページは抽出対象から除く．その上でキーワード a の出現区間中で a が最後に出現しているページと, キーワード b の

出現区間中で b が最初に出現しているページの抽出を行い，補完ページとしてユーザに提示する．例えば，あるページで「森総理」と「小泉総理」というキーワードの時系列パターンを求めたとき,「森総理」が時間的に前に,「小泉総理」が時間的に後に出現する前後関係にあった場合，時系列ページ中でキーワード「森総理」が出現する最後のページ，キーワード「小泉総理」が出現する最初のページを抽出する．

5.2 出力例の提示

実際にサンプルデータを用いて，各パターンにおいて出力されるページを示す．サンプルは実データをもとにして作成したものである．

5.2.1 共起的反復の出力例

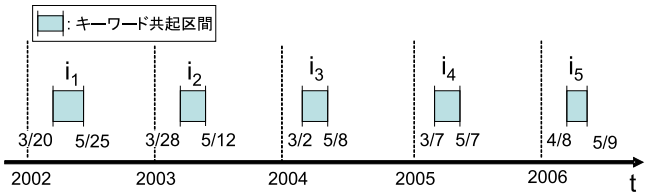

図 4　共起的反復 区間の抽出

共起的反復と判定されたページのサンプルとして図 4 のデータを用いてスコアの計算例を示す．図 4 は一つの時系列ページ中のキーワードの出現状況を表しており，図 4 中の長方形が，時系列ページ中のキーワードの共起区間を表している．キーワード a は「桜」，キーワード b は「花見」である．また，現在のページでは「桜」と「花見」が共に出現しているとする．各区間のスコアの計算結果を表 1 に示す．

表 1　共起的反復 区間のスコアリング

	区間の長さ	区間の長さの平均値	各区間のスコア
i_1	66.0	55.6	0.82
i_2	45.0		0.81
i_3	67.0		0.80
i_4	69.0		0.76
i_5	31.0		0.56

次に，ページのスコア計算の例を示すため，キーワード共起区間 i_2 の詳細を図 5 に示す．図中の長方形が区間 i_2 中に存在するページを表しており，文字はページ中に出現しているキーワードを表している．共起区間 i_2 中のページを時間的に前のものから順に $p1, p2, \cdots, p8$ とする．$score(a)$，$score(b)$ をそれぞれ計算し，ページごとのスコアを算出する．ここでは現在のページにキーワード a,b が共に出現している場合を想定しているため，$score(a), score(b)$ に対する補正は行われない．

各ページのスコアを算出した後，各ページのスコアに対し区間が持つスコアを積算し最終的なページのスコア score($page$) を算出する．i_2 中の各ページのスコアを表 2 に示す．

5.2.2 排他的反復の出力例

排他的反復と判定されたページのサンプルとして図 6 のデータを用いてスコアの計算例を示す．図 6 は一つの時系列ページ中のキーワードの出現状況を表しており，図 6 中の長方形が，

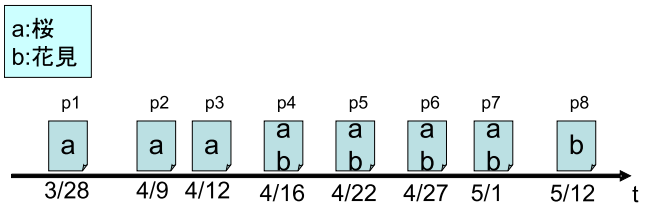

図 5　共起的反復 ページの抽出

表 2　共起的反復 ページのスコアリング

	$score(a)$	$score(b)$	$score(a) + score(b)$	$score(page)$
$p1$	0.02	-	0.02	0.01
$p2$	0.35	-	0.35	0.28
$p3$	0.44	-	0.44	0.35
$p4$	0.55	0.04	0.59	0.48
$p5$	0.73	0.23	0.96	0.78
$p6$	0.88	0.40	1.28	1.03
$p7$	1.0	0.57	1.57	1.27
$p8$	-	1.0	1.0	0.81

時系列ページ中のキーワードの出現区間を表している．キーワード a は「桜」，b は「紅葉」である．サンプルではキーワード「桜」とキーワード「紅葉」が交互に出現しており，現在のページには「紅葉」のみが出現しているとする．条件より，抽出対象となるのはキーワード a の出現区間となり表 3 中の i_1, i_2, i_3 が補完ページの抽出対象区間となる．抽出対象区間におけるスコアの計算結果を表 3 に示す．

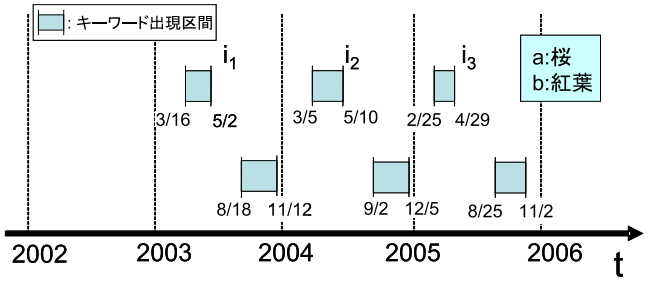

図 6　排他的反復 区間の抽出

表 3　排他的反復 区間のスコアリング

	区間の長さ	区間の長さの平均値	区間のスコア
i_1	47.0	58.7	0.80
i_2	66.0		0.88
i_3	63.0		0.93

ページのスコア計算の例を示すため，キーワード出現区間 i_3 の詳細を図 7 に示す．図中の長方形が区間 i_3 中に存在するページを表しており，文字はページ中に出現しているキーワードを表している．i_3 中のページを時間的に前のものから順に $p1$, $p2$, $p3$, $p4$, $p5$ とし，各ページのスコアを算出する．さらに，各ページのスコアに区間の持つスコアを積算し，最終的なスコア score($page$) を算出する．スコアの計算結果を表 4 に示す．

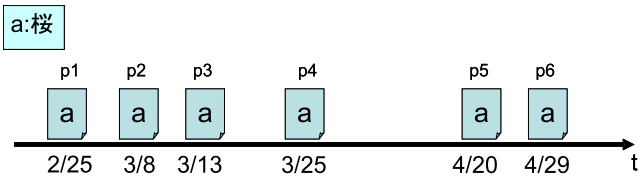

図 7　排他的反復 ページの抽出

表 4　排他的反復 ページのスコアリング

	$score(a)$	$score(page)$
$p1$	0.02	0.01
$p2$	0.17	0.16
$p3$	0.25	0.23
$p4$	0.44	0.40
$p5$	0.85	0.79
$p6$	1.00	0.93

5.2.3　前後関係の出力例

キーワードが前後の時系列パターンを持つページのサンプルとして図 8 に示すデータを用いる．サンプルはキーワード「森総理」が時間的に前に出現し，キーワード「小泉総理」が時間的に後に出現する関係にある．前後では時間的に先に出現しているキーワードの最後に出現しているページと，時間的に後に出現しているキーワードの最初に出現しているページを補完ページとして取得する．キーワードの出現区間に重複がある場合は，重複している区間を除き，その上でページを決定する．

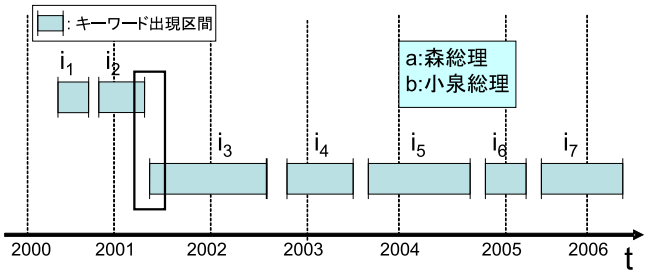

図 8　前後 区間の抽出

図 8 中の黒い枠で囲われている部分の詳細を図 9 に示す．図 9 中の太い枠で囲まれているページが抽出ページとなる．

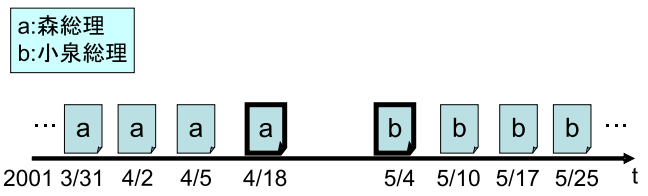

図 9　前後 区間詳細

6.　プロトタイプシステムと評価実験

6.1　プロトタイプシステムの構築

本節では，プロトタイプシステムの構成について述べる．プロトタイプシステムは以下によって構成される．

(1)URL 収集部　ユーザの入力した質問キーワードから質問を生成する．質問キーワードが a, b, c の場合には，$a, b, c, a \wedge b, b \wedge c, a \wedge c, a \wedge b \wedge c$ が生成される．質問生成部によって生成された質問によって Web ページを検索し，検索結果として得られた Web ページの URL を収集する．このとき，検索エンジンには Google を利用する．

(2) 時系列ページ収集部　収集された URL をキーとして，Web アーカイブから過去のページを自動的に取得する．過去のページの取得はアメリカの InternetArchive(注3) を用いて行う．InternetArchive ではページの URL を入力することで，過去のページを取得することが出来る．このとき，InternetArchive 内の情報から，ページが InternetArchive に収集された年月日をページの時間情報として取得する．

(3) キーワード出現時間抽出部　収集したページのソースコードをテキスト情報として保存する．その際，ページの URL と時間情報をインデクスとしてテキスト情報と共に保存する．保存されている全てのページのテキスト情報に対して全文検索を行い，キーワードが出現しているページの時間情報を検索結果として返す．検索には Interstage Shunsaku Data Manager(注4) を使用する．

(4) 時間的関係判定部　キーワードの出現しているページの時間情報をもとに，各 Web ページにおけるキーワードの出現パターンを解析し，時系列パターンを判定する．

(5) 補完ページ抽出部　判定されたキーワードの時系列パターンに基づき，同一 URL の Web ページの過去のページから現在の内容を補完するページを抽出する．

6.2　評 価 実 験

補完ページの抽出方式に対する評価を行うため実験を行った．まず URL を任意に 150 件収集し，それらの過去のページを取得して実験対象とした．収集したページに対して質問キーワードの出現状況を調べ，質問キーワードを全て含んでいるページに対して時系列パターンの判定を行い，判定結果に基づいて補完ページの抽出を行った．時系列パターンの判定に用いる閾値は以下の通り設定し，キーワードが共起しているとみなせる範囲は，キーワードが出現しているページ間の時間的距離が 30 日以内であるものとした．

- α = 0.5
- β = 0.4
- γ = 5
- δ = 0.3
- p = 5
- q = 3

質問キーワードは，時系列パターンの性質を考えパターンが発生しやすいと思われるものを任意に選んだ．

6.3　実 験 結 果

結果として，時系列パターンが共起的反復であるページが 3 件，排他的反復であるページが 4 件，前後関係であるページが

(注3)：http://www.archive.org/
(注4)：http://interstage.fujitsu.com/jp/shunsaku/

2 件抽出された．判定を行ったキーワードとページの URL，判定された時系列パターンについて表 5 に示す．以下に，実際に抽出された補完ページの例を挙げる．

(実験番号 1) 質問キーワード { 桜，花見，観光 }

実験対象のうち，キーワードを時系列的に全て含む URL は 3 件であった．そのうちの 2 件では全てのキーワードの組み合わせが時間的非依存となった．残りの 1 件では，{ 桜，花見 } は共起的反復，{ 桜，観光 }，{ 花見，観光 } は時間的非依存となった．この URL では地域の観光情報を紹介しており，時期に合わせたイベント情報を掲載していた．実際にページ内容を確認したところ，毎年春に「桜」と「花見」が出現していたことから，{ 桜，花見 } が共起的反復と判定されたのは妥当であるといえる．

この URL の現在のページと，補完ページとして抽出された過去のページのイメージを図 10 に示す．キーワード「観光」が出現している現在のページに対し，補完ページとしてキーワード「桜」,「花見」が出現しているページが提示されている．キーワードの全共起区間に含まれるページ数は 32 ページであり，補完ページとして抽出されたのは 9 ページであった．いずれも桜の開花や花見の見所の情報を掲載しているページが抽出された．

図 10　共起的反復 補完例

(実験番号 7) 質問キーワード { 節分，ひなまつり }

実験対象のうち，キーワードを時系列的に全て含む URL は 6 件であった．そのうちの 5 件で { 節分，ひなまつり } は時間的非依存となった．残りの 1 件では { 節分，ひなまつり } は排他的反復と判定された．この URL はコンビニの商品情報を掲載しており，イベントにあわせた商品についての情報が掲載されていた．節分の時期には節分用の巻き寿司の商品情報が，ひなまつりの時期にはひなまつり用のケーキの商品情報が掲載されており，排他的反復と判定されたのは妥当であるといえる．

このページの現在のページと，補完ページとして抽出された過去のページのイメージを図 11 に示す．キーワード「ひなまつり」が出現している現在のページに対し，補完ページとしてキーワード「節分」が含まれる過去のページが提示されている．キーワード「節分」の全出現区間に含まれるページ数は 33 ページであり，補完ページとして抽出されたのは 9 ページであった．いずれも「節分」のイベント商品について掲載されているページが抽出された．

図 11　排他的反復 補完例

(実験番号 8) 質問キーワード { 近鉄，オリックス }

実験対象のうち，キーワードを時系列的に全て含む URL は 5 件であった．そのうちの 4 件で { 近鉄，オリックス } は時間的非依存と判定された．残りの 1 件では { 近鉄，オリックス } は前後関係と判定された．この URL は，元はプロ野球チームの近鉄バファローズのホームページであり，同じプロ野球チームであるオリックスブルーウェーブとの合併後，そのまま使用されているものであることから，前後関係の判定は妥当であるといえる．

前後関係と判定されたページの現在のページと補完ページとして抽出された過去のページのイメージを図 12 に示す．現在のページ内容は，プロ野球チーム近鉄バファローズとオリックスブルーウェーブの合併後のチームについての現在の情報であり，補完ページとしてチームが合併する前後の時期のページが抽出された．

図 12　前後 補完例

6.4　考　　察

時系列パターンの判定においては，ほとんどのページが時間的非依存と判定された．主な原因として，キーワードがページ中に常に出現していた場合は，全て時間的非依存と判定されてしまうということが挙げられる．本方式では，時系列ページにおけるキーワードの出現パターンの判定はページ中にキーワードが出現しているかどうかのみに依存しており，キーワードのページ中での出現数の変化やページの更新の有無などは考慮し

表5 実験結果

実験番号	質問キーワード	URL	時系列パターン
1	{ 桜，花見，観光 }	$http://www.biwako-visitors.jp/$	{ 桜，花見 }：共起的反復 { 花見，観光 }：時間的非依存 { 観光，桜 }：時間的非依存
2	{ 共同通信杯，シルクロード S，競馬 }	$http://www.netkeiba.com/$	{ 共同通信杯，シルクロード S}：共起的反復 { シルクロード S，競馬 }:時間的非依存 { 競馬，共同通信杯：時間的非依存 }
3	{ 京都記念，札幌記念，競馬 }	$http://keiba.yahoo.co.jp/$	{ 京都記念，札幌記念 }：排他的反復 { 札幌記念，競馬 }：時間的非依存 { 競馬，京都記念 }：時間的非依存
4	{ 伊勢丹，おせち }	$http://www.isetan.co.jp/icm2/jsp/shops/foods/oseti/index.jsp$	{ 伊勢丹，おせち }：共起的反復
5	{ お中元，お歳暮 }	$http://www.isetan.co.jp/icm2/jsp/shopping/seasongift/top.ac$	{ お中元，お歳暮 }：排他的反復
6	{ お中元，お歳暮 }	$http://www.e-meitetsu.com/e-shop/gift/$	{ お中元，お歳暮 }：排他的反復
7	{ 節分，ひなまつり }	$http://www.lawson.co.jp/go_lawson/index.html$	{ 節分，ひなまつり }：排他的反復
8	{ 近鉄，オリックス }	$http://www.buffaloes.co.jp/$	{ 近鉄，オリックス }：前後
9	{ 森総理，小泉総理 }	$http://www.kantei.go.jp/$	{ 森総理，小泉総理 }：前後

ていない．よって，キーワードが常に出現している場合，キーワードの出現の周期を判定できず時間的非依存と判定される．今後はページ中のキーワード数の変化や，ページの更新などを考慮した判定方法を考えていく必要がある．

補完ページの抽出方式については，補完ページとして現在のページに存在しないキーワードに関する情報を含んだページが抽出されており，方式は有効であると考えられる．しかしながら，補完ページを抽出する際に一つのキーワード出現区間から偏ってページが抽出される場合があり，その場合，抽出された複数の補完ページの内容がほぼ同一のものになってしまうことがあった．

7. まとめと今後の課題

本稿では，質問の緩和によってキーワードを時系列的に含んでいる Web ページを抽出し，ページの過去の内容をキーワードの出現パターンに基づいて抽出することで，同一 Web ページの過去のページにより現在のページを補完してユーザに提示する方式について提案した．さらに，実際に過去のページを収集し，補完ページの抽出についての実験を行った．今後の課題として，時系列パターンの判定アルゴリズムの見直し，補完ページ抽出方式の改良などが挙げられる．

謝　　辞

本研究の一部は，平成 16 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．

文　　献

[1] Adam, J., 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).

[2] Adam, J., 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).

[3] 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).

[4] Kage, T. and Sumiya, K.: A Web Search Method Based on the Temporal Relation of Query Keywords, *Proc. of The 7th International Conference on Web Information Systems Engineering (WISE 2006)*, pp. 4–15 (2006).

[5] Yumoto, T. and Tanaka, K.: Finding Pertinent Page-Pairs from Web Search Results, *Proceedings of The 8th International Conference on Asian Digital Libraries (ICADL2005)*, pp. 301 – 310 (2005).

[6] Yumoto, T. and Tanaka, K.: Page Sets as Web Search Answers, *Proceedings of The 9th International Conference on Asian Digital Libraries (ICADL2006)*, pp. 244 – 253 (2006).

[7] 池田新平，是津耕司，小山　聡，田中克己：Web コンテンツの周辺情報提示によるナビゲーション支援，日本データベース学会論文誌 Vol.2 No.1，日本データベース学会，pp. 139–142 (2003).

[8] 郡　宏志，竹原幹人，大島裕明，小山　聡，田中克己：Blog Radio Blog 情報の感情マイニングと可聴化に基づく Web 閲覧補完，第 16 回データ工学ワークショップ (DEWS2005) 論文集，電子情報通信学会 (2005).