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

賀家　智代†　　角谷　和俊††

† 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: †nd05w005@stshse.u-hyogo.ac.jp, ††sumiya@shse.u-hyogo.ac.jp

あらまし　近年，Web ページを収集・保存する Web アーカイブの構築が進められている．しかしながら，Web アーカイブを有効利用し，そこから利用者にとって有益な情報を効率よく取得する方法についての検討はほとんどなされていない．本研究では，Web アーカイブの時系列特性を考慮した新たな検索方式を提案する．本方式は 2 つの機構で構成される．1 つはユーザが入力した質問キーワードの順序関係を判定する機構で，アーカイブに格納されている時系列の Web ページにおけるキーワードの出現パターンを解析することによってキーワードの順序関係を判定する．もう 1 つはその順序関係に基づいて質問を生成し検索する機構で，各々の順序関係に対応する順序関係を組み合わせて質問を自動生成する．本稿では，提案する 2 つの機構について述べ，そのプロトタイプの設計および順序関係判定方式の評価実験を検討する．

キーワード　Web アーカイブ，情報検索

A Web Archive Search Engine Based on the Temporal Relation of Query Keywords

Tomoyo KAGE† and Kazutoshi SUMIYA††

† 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: †nd05w005@stshse.u-hyogo.ac.jp, ††sumiya@shse.u-hyogo.ac.jp

Abstract　Web pages are being collected and stored in Web archives, and several methods to construct Web archives have been developed. Although there are few studies using Web archives effectively, and it is hardly considered how to get useful information for a user. We propose a method to retrieve time series of Web pages from Web archives by using the pages' temporal characteristics. We present two processes for searching Web archives based on the temporal relation of query keywords. One is a method for determining the temporal relation of query keywords. The other is a method of inquiry based on the relation. In this paper, we discuss the two processes and have considered experiment and implementation issues regarding a prototype system.

Key words　Web Archive, Information retrieval

1. はじめに

インターネット上で公開されている Web コンテンツは，更新・削除が容易なため頻繁に内容が改変されたりページが消失したりする．そこで，Web ページを収集し永久に保存する Web アーカイブの構築が各国で行われ，効率的なクローリング技術や永久的に Web ページを保存する技術 [1], [2]，フォーマットリポジトリに関する研究 [3]，コンテンツ間の一貫性を保ちコンテンツの同一性を保障するための方式 [4] など多くの構築手法が検討されている．

また，大規模データマイニング [5], [6], [7] や Web アーカイブの要約に関する研究 [8], [9], [10], [11] など Web アーカイブに格納されている大規模な時系列データを活用する研究も行われている．しかしながら，Web アーカイブから情報を取得する試みはほとんどなされておらず，現在提案されている Web における検索方式 [12], [13] を Web アーカイブに適用しても，出力が

アーカイブされた量と比例して膨大となり利用価値は少ない．

そこで我々は，Web ページによってキーワードの捉え方が異なる点に着目し，時系列データにおけるキーワードの出現パターンを利用した Web アーカイブの検索方式を提案する．本研究でのキーワードの出現パターンとは，複数のキーワードが時系列に出現する順序依存関係を指し，そのパターンによってキーワード同士の関係を判定する．我々は,「ある Web ページにおけるキーワードの関係」=「ある Web ページにおけるキーワードの捉え方」と考え，等しい捉え方の Web ページをまとめてユーザに出力することにより膨大な Web ページの中から意図する情報を見つけやすくすることを支援する．

本方式は，ユーザによって複数のキーワードが入力されると，そのキーワードの出現パターンを解析し，同じ順序関係のページをクラスタリングして呈示する．また，1 つの URL 内でもキーワードの出現状況に基づき同じ内容のページをまとめて出力する．機構は以下の通りである．

- ユーザによって入力されたキーワードの順序関係の判定
- 順序関係に基づくクラスタリングと質問生成

本稿では上記手法に加え，その実験とプロトタイプシステムについて検討する．構成は次の通りである．2 節では動機と本研究の概要について述べ，3 節では質問キーワードの順序関係の判定方法について説明する．4 節では 3 節で得られた順序関係に基づく質問生成法について述べ，5 節では提案する方式に基づいた実験とプロトタイプシステムの設計について述べる．最後に 6 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

Web アーカイブに格納された Web ページは，削除されて存在しない情報だけでなく，時間と共に変化する情報 [14] や，ある時間に依存した情報が格納されており，様々な観点での検索が可能であると考えられる．従来のキーワードを質問とする Web 検索ではそのキーワードを含む Web ページが検索結果となり，複数のキーワードによる検索では一般に AND 条件が用いられる．しかし，異なる時間に出現するキーワード，例えば「花見」と「紅葉狩り」といった場合には，意図した出力結果が得られにくい．

そこで本研究では，Web アーカイブに格納されている時系列ページを検索対象としてユーザが質問キーワードにより問い合わせを行った場合に，質問キーワードの持つ時間的意味を考慮した検索を提案する．本方式はキーワードの有無ではなく複数のキーワードの時間的順序関係，すなわち出現状況に基づく．キーワードの順序関係を用いることによって,「花見」と「紅葉狩り」が同一ページに存在しない場合でも取得することができる．概要を図 1 に示す．

まず，時間的関係を判定するために，ユーザが入力した質問キーワードを 2 つずつの組に分解する．次に，順序関係が保たれている区間を計算し，その区間を用いてキーワードの順序関係を判定する．このとき，2 つのキーワードが同時期に出現する共起関係 (co-occurring)，あるキーワードが他のキーワード

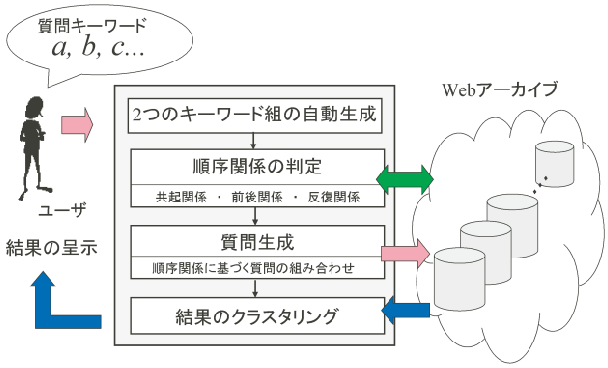

図 1 概　　要

よりも前に出現するという前後関係 (ordered)，あるいはあるキーワードと他のキーワードが交互に出現するという反復関係 (repeated) の 3 つの関係に分類する．ここで，同一キーワードであっても組み合わせにより相手のキーワードが異なる場合は，順序関係が変化することに注意を要する．

さらに，順序関係から質問を生成し，その質問を組み合わせて複数の質問を再構築する．再構築した質問によって，キーワードの関係別に Web ページをクラスタリングし検索結果とする．検索結果は，URL を順序関係毎にをクラスタリングすることによって順序関係の等しい URL をまとめ，キーワードの順序関係が等しい URL 毎に呈示する．

2.2 関 連 研 究

Web アーカイブの検索エンジンとして，WayBack Machine(注1)や WERA(注2)が提案されている．しかしながら，前者は URL の入力と時間の指定が必要なため意図した情報を取得することは難しい．また，後者はキーワードによる検索が可能であるが，全文検索の機能が提供されているのみで Web アーカイブの特性が考慮されていない．また，時系列 Web ページのための検索エンジンとして奥村らの blogWatcher(注3)があるが，対象が blog に特化されている点，キーワードの絶対時間について分析されている点で本研究と異なる．

Web アーカイブからの情報取得に関する研究として，Web アーカイブの要約方式 [8] [9] [10] [11] や過去のコンテンツを考慮した検索結果の再ランキング方式 [15] [16] が Adam らによって提案されている．キーワードをトピックとして扱い，更新による内容の差分を抽出するのに対して，本研究では，時系列のページ全体の傾向に基づきキーワード間の関係を分析し，過去の Web ページを検索する点で，これらの研究と異なる．

質問キーワードの時系列性を扱った研究として，Chien らによる [17] や Vachos らによる [18] 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 ページに出現するキーワードの傾向を解析して各 URL の傾向を抽出し，その URL から適合する Web ページを検索することを目的としている．

(注1)：http://www.archive.org/
(注2)：http://archive-access.sourceforge.net/projects/wera/
(注3)：http://blogwatcher.pi.titech.ac.jp/

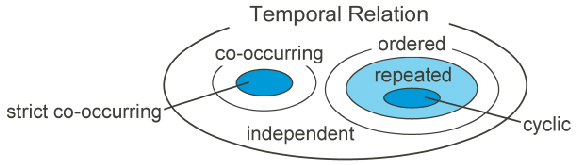

図 2　キーワードの順序関係

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

我々は，ユーザが Web アーカイブ検索エンジンに入力した質問キーワードの関係について 6 つの順序関係を定義する．キーワード間の順序関係を図 2 に示し，以下に述べる．なお，順序依存関係を最初に述べる共起，前後，反復とし，それに非依存関係を合わせて基本的な順序関係とする．基本形の強い関係として「共起」の強い関係を「密共起」,「反復」の強い関係を「循環」とする．

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関係を共起関係という．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，本研究では同一 URL の時間が異なる複数のページを対象としているため共起の範囲をページとはしない．すなわち，共起の範囲はページ単位ではなく時間単位である．そのため，キーワード a, b が 1 つのページに出現していなくてもおおよそ同じ時期に出現する場合は共起関係があるといえる．

前後（ordered） キーワード a, b が互いに順序依存している関係，すなわち時系列ページにおいて一方が他方より常に時間的に前に出現する関係を前後関係という．前後関係である 2 つのキーワードは，一方が他方より常に前または後に出現し，ノイズを除いて順序が逆になることはない．例えば，試験制度の改定によって資格の名称が変更された「第二種情報処理」と「ソフトウェア開発」が挙げられる．例でも分かるように前後関係は因果関係を包含している．後述する反復については繰り返す前後関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関係，すなわち a と b が時系列ページに交互に出現する関係を反復関係という．例えば,「応募」と「当選」が挙げられる．反復関係は 1 つの順序関係では決定されず複数の前後関係が成立するため前後関係に包含される．特徴としては，前後関係が 1 つの順序関係しかないのに対して反復関係は 2 つの順序（前後）関係がある点である．つまりキーワードの半順序関係が満たされている時区間が 2 種類（「a が先で b が後」という区間と「b が先で a が後」という区間）存在するということである．後述する循環については厳密な反復関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関係，すなわちキーワードが互いにランダムに出現する関係を非依存関係という．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．これらのキーワードは時間の流れに依存しないため一般的に例えることは難しい．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．

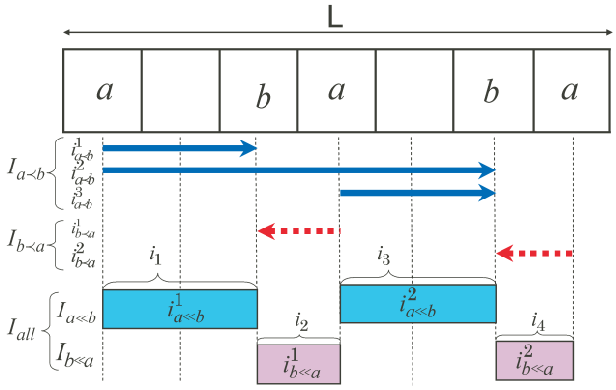

図 3　キーワードの関係

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，時系列ページにおいて，キーワード a, b の大部分が共に時系列ページの 1 時点のページに出現している関係をいう．例えば,「放射線」と「土壌」は密共起になる．

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．具体的には，a から b の時区間が常に一定で，同様に b から a の時区間も等しい場合をいう．例えば「春」と「秋」という 2 つのキーワードの出現間隔は，毎年「春」から「秋」が一定で，同様に「秋」から「春」も等しいため循環関係といえる．また，社会的なイベント等は規則的に行われているため，それらのキーワードは循環関係になる可能性が高いと考えられる．

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．例えば，質問キーワードが a, b, c の 3 つの場合，$\{a, b\}$, $\{b, c\}$, $\{c, a\}$ の 3 つの組が自動生成され，それぞれの組について判定される．キーワードの組は入力したキーワード数が n 個の場合，${}_nC_2$ 通り生成される．生成した全てのキーワードの組について時区間を抽出し，その時区間を演算して 3.1 節で述べた順序関係に分類する．判定する任意のキーワードを a, b として，定義を図 3 に示し以下に述べる．

3.2.1　順序関係判定のための時区間抽出

順序関係を判定するために時区間を抽出する．まず，$i_{a\prec b}$ と $i_{b\prec a}$ を抽出する．$i_{a\prec b}$ は a を含むページを始端，b を含むページを終端とする時区間で，$I_{a\prec b}$ はその集合を表す．$i_{b\prec a}$ は b を含むページを始端，a を含むページを終端とする時区間で，$I_{b\prec a}$ はその集合を表す．次に，それらの区間を演算して $i_{a\ll b}$ と $i_{b\ll a}$ を抽出する．$i_{a\ll b}$ は a が b よりも先に出現する区間で,$I_{a\ll b}$ はその集合を表す．$i_{b\ll a}$ は b が a よりも先に出現する区間で,$I_{b\ll a}$ はその集合を表す．$i_{a\ll b}$ は，まず，$I_{a\prec b}(I_{a\prec b} = \{i^1_{a\prec b}, i^2_{a\prec b}, ..., i^n_{a\prec b}\})$ を抽出し，$i_{b\prec a}$ を含んだ $i_{a\prec b}$ を取り除くことで抽出することができる．

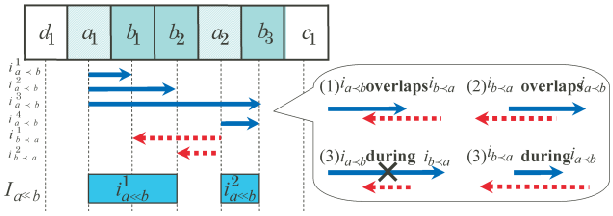

図 4　時区間抽出の例

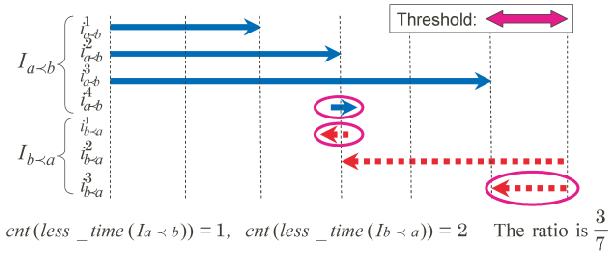

図 5　共起の例

時系列ページにおける時区間を次のように定義する．L は時系列ページの実時間の長さである．$I_{all}(=\{i_1, i_2, ..., i_n\})$ は $I_{a\ll b}$ と $I_{b\ll a}$ の集合を表す．

関数を次のように定義する．$time(i)$ は時区間 i の実時間を返す．$reverse(i_1, i_2)$ は真か偽かを返す．もし $((i_1 == i_{a\ll b}) \wedge (i_2 == i_{b\ll a}))$ または $((i_1 == i_{b\ll a}) \wedge (i_2 == i_{a\ll b}))$ ならば $reverse(i_1, i_2)$ は真を返す．そうでなければ偽を返す．$variance(x)$ は x の分散値を返す．$cnt(I)$ は I に含まれている時区間の総数を返す．$time\_0(i)$ は実時間 0 の i を返す．$less\_time(I)$ は I に含まれている閾値以下の時区間の総数を返す．$\alpha, \beta, \gamma, \theta$ 及び δ は閾値を表す．

3.2.2　共起関係の判定

時系列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大きいと共起関係と判定する．同時期といえる時区間の閾値を定め，キーワード a, b が閾値時間以内に出現している場合を共起とする．共起の抽出にはキーワードの出現間隔と見なせる $i_{a\prec b}$ と $i_{b\prec a}$ を用いる．$time(i_{a\prec b})$ により出現時間間隔を求め，この値が閾値以下である $i_{a\prec b}$ と $i_{b\prec a}$ の総数によって共起数が求められる．そして全体の出現数に占めるその総数の割合が閾値 α より大きければ共起関係とみなす．共起関係であると判定された場合，同一ページに出現する場合も考慮し，$time(i_{a\prec b}) = 0$ のみの値の割合が閾値 α を上回る場合を密共起関係とみなす．

```
//共起関係の判定
if ( (cnt(less_time(I_{a≺b}))+cnt(less_time(I_{b≺a})))/(cnt(I_{a≺b})+cnt(I_{b≺a})) ≧ α)
    then //循環関係の判定
        if ( cnt(time_0(i_{a≺b}))/(cnt(I_{a≺b})+cnt(I_{b≺a})) ≧ α)
            then
                strict co-occurring
            else
                co-occurring
        fi
fi
```

3.2.3　前後関係の判定

まず，時系列ページにおいて順序関係が全体的に成立しているか否かを判定するために I_{all} が時系列ページ全ての時間 (L) に占める割合を計算する．I_{all} は順序関係が保たれている区間（$I_{a\ll b}$ や $I_{b\ll a}$）なので，順序が保たれている区間が時系列ページ全区間に占める割合が大きいとキーワードは順序依存していると考えられる．この値が閾値 β より小さければ終了，大きければ次のステップに進む．次に，$i_{a\ll b}$ の和が L に占める割合と $i_{b\ll a}$ の和が L に占める割合を計算する．この値の偏りが閾値 γ 以上であれば前後関係とみなす．

```
//前後関係の判定
if ( (Σ_{i=1}^{n} (time(i_i)))/L ≧ β )
    if ( (Σ_{i=1}^{l} (time(i^i_{a≪b})))/(Σ_{i=1}^{n} (time(i^i_{b≪a}))) ≧ γ)
        then
            ordered a < b
    fi
    if ( (Σ_{i=1}^{n} (time(i^i_{b≪a})))/(Σ_{i=1}^{l} (time(i^i_{a≪b}))) ≧ γ)
        then
            ordered b < a
    fi
fi
```

3.2.4　反復関係の判定

全体的に反復している区間が成立してれば反復関係とする．したがって，I_{all} が L に占める割合が閾値 β より小さければ終了，大きければ次のステップに進む．次に，順序が保たれている区間 $i_{a\ll b}$ と $i_{b\ll a}$ が交互に出現しているかを判定する．交互であれば区間を足し合わせ，I_{all} 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．反復関係である場合，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 $i_{a\ll b}$ の各々の時区間が一定であるので，（周期と見なせる）i_i 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(注4)．

```
//反復関係の判定
if ( (Σ_{i=1}^{n} (time(i_i)))/L ≧ β )
    then
        foreach (i_j)
            if (reverse(i_j, i_{j+1}) == true)
                s = Add(time(i_j));
                j++;
            fi
        end
        if ( s/(Σ_{i=1}^{n} (time(i_i))) ≧ θ )
            then //循環関係の判定
                if ((variance(i_{a≪b})) < δ) ∧ (variance(i_{b≪a})) < δ)
                    then
                        cyclic
                    else
                        repeated
                fi
        fi
fi
```

(注4)：分散値はデータの散ばり具合を表す指標である

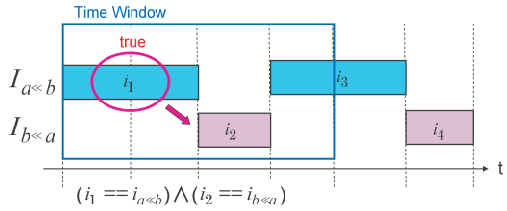

図 6　反復関係の例

表 1　生成される質問と抽出区間

関係	質問	抽出される区間
共起	$Q_{a \oplus b}$	$I_{\{a \wedge b\}}$
密共起	$Q_{a \otimes b}$	$I_{\{a \wedge b\}}$, Keywords a and b should appear in the same page.
前後	$Q_{a<b}$	$I_{a \ll b}$
反復	$Q_{a \circ b}$	$I_{a \ll b} \cup I_{b \ll a}$
循環	$Q_{a \odot b}$	$I_{a \ll b} \cup I_{b \ll a}$, Each period should be the same time length.
非依存	$Q_{\{a,b\}}$	$I_{\{a \vee b\}}$

図 6 は反復関係判定方式の例を示している．図に示すように，time window は長すぎる区間をカウントしない役割を果たしている．この場合，時系列に隣接した時区間は $\{i_1, i_2\}$, $\{i_2, i_3\}$ 及び $\{i_3, i_4\}$ である．i_1 に注目すると，i_1 は $i_{a \ll b}$ で，その隣の i_2 は $i_{b \ll a}$ であるので i_1 は反復していると考えられる．

3.2.5　非依存関係の判定

以上の共起，前後，反復関係に該当しない場合，非依存関係とみなす．前述した 3 つの判定方法は互いに排他的であるため重複して判定されることはない．つまり順序依存していれば一意に決まり，順序依存していない場合は非依存関係となる．

4.　順序関係に基づく質問生成

4.1　質 問 生 成

質問の要素 Q によって返される時区間（各々の時区間は時系列ページを含んでいる）を表 1 に示す．

$I_{\{a \wedge b\}}$ は，a と b が時間的に近隣に出現している区間の集合を表す(注5)．$I_{\{a \vee b\}}$ は，a または b のどちらかが含まれているページの時区間の集合を表す．

$Q_{a \otimes b}$ は密共起関係で，a と b の両方を含んでいるページの時区間を返す．$Q_{a \odot b}$ は循環関係で，平均周期で分割した時区間を返す．$Q_{a \oplus b}$ と $Q_{\{a,b\}}$ はキーワードの出現状況について問い合わせ，$Q_{a<b}$, $Q_{b<a}$ 及び $Q_{a \circ b}$ はキーワードの順序に基づく時区間を問い合わせる．このような異なる特性の質問要素を組み合わせることによって質問生成する．

4.2　質問の組み合わせ

ユーザが入力した全てのキーワードを含むように質問を組み合わせて質問を生成する．

例 1　質問キーワード a, b, c における順序関係が $\{a, b\}$ は共起，$\{b, c\}$ と $\{c, a\}$ が非依存であるとする．質問の要素は $Q_{a \oplus b}$, $Q_{\{b,c\}}$ 及び $Q_{\{c,a\}}$ で，キーワード a と b は互いに順序依存し，

(注5)：近隣の範囲はアプリケーションに依存する．

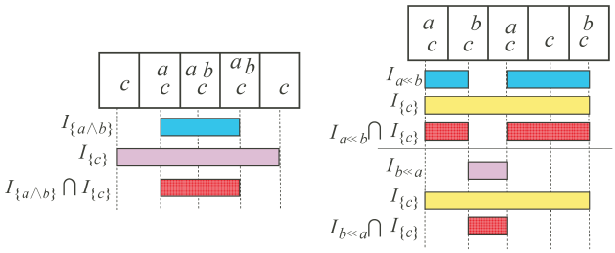

図 7　質問の組み合わせ

c は独立となるため，全てのキーワードを含むように質問の組み合わせると $Q_{a \oplus b} \wedge Q_{\{c\}}$ となり，抽出される区間は以下のようになる（図 7 の左部分）．

$$R(Q_{a \oplus b} \wedge Q_{\{c\}}) = I_{\{a \wedge b\}} \cap I_{\{c\}}$$

例えば，a が“ SARS ”，b が“ corona ”，c が“ virus ”という組み合わせが考えられる．

例 2　次の例は，$\{a, b\}$ が反復，$\{b, c\}$ と $\{c, a\}$ が非依存である場合を想定する．質問の要素は，$Q_{a \circ c}$, $Q_{\{a,b\}}$ 及び $Q_{\{c,b\}}$ となり，全てのキーワードを含むように質問の組み合わせると $Q_{a \circ b} \wedge Q_{\{c\}}$ となり，抽出される区間は以下のようになる（図 7 の右部分）．

$$\begin{aligned} R(Q_{a \circ b} \wedge Q_{\{c\}}) &= (I_{a \ll b} \cup I_{b \ll a}) \cap I_{\{c\}} \\ &= (I_{a \ll b} \cap I_{\{c\}}) \cup (I_{b \ll a} \cap I_{\{c\}}) \end{aligned}$$

この場合，出力される区間は $I_{a \ll b}$ または $I_{b \ll a}$ と $I_{\{c\}}$ が重複した部分である．図 7 の右部分の上部は $I_{a \ll b} \cap I_{\{c\}}$ を，下部は $I_{b \ll a} \cap I_{\{c\}}$ を表している．例えば，a が「お中元」，b が「お歳暮」，c が「ギフト」という組み合わせが考えられる．

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 8 に示す．なお，太枠及び太線部分は本システムの特有の機構で，以下のような 5 つのユニットから成る．

（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，その URL を指定することによって時系列の Web ページを収集する．取得方法は，InternetArchive(http://web.archive.org/web/*/任意の URL) から，そこに張られたリンクによって抽出する．ただし，アドレスの変化を考慮して 1 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．さらに，エラーページを除くためにサイズが 5byte 未満の Web ページの除去する．Internet Archive の典型的なエラーページ（コンテンツが存在しないページ）が 5byte 弱であるため，この値を用いる．

（2）　**インデックス生成部**　ページのテキストを形態素解析し，名詞を抽出する．名詞とページが収集された時間をページのインデックスとして記述する．時間データは，InternetArchive(http://web.archive.org/web/14 桁の数値/任意の URL) の数値部分から取得する(注6)．

(注6)：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．

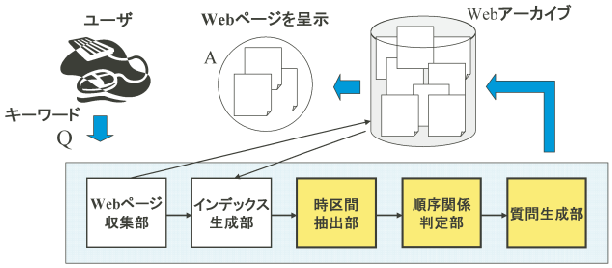

図 8　システム構成図

（3） **時区間抽出部** 順序関係を判定するための時区間を抽出する．時区間の抽出は，インデックスを参照し，各 URL 毎に行う．

（4） **順序関係判定部** 時区間を演算し，2 つのキーワードの順序関係を判定する．

（5） **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．

5.2 実　　験

順序関係の判定方式及びそれに基づく 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．Web ページ収集部により約 100 個の URL を時系列 Web ページを自動的に取得し，質問は 3 つのキーワードで構成し，各々の質問を複数の Web ページ (URL) に対して問い合わせた．質問に対する URL の選定基準は質問キーワードを含んでいるものを任意に選んだ．閾値は重複して順序関係が出ず，非依存関係が多くならないように事前に調査して α, β, γ をそれぞれ 0.3, 0.6,5，共起の範囲を 90 日としてキーワードの順序関係の判定を行った．

5.2.1 実　　験 1

判定された順序関係から生成される質問によって，1 つの URL の時系列ページ群をクラスタリングする実験を行った．

[実験 1.1] {SARS, corona, virus} を質問キーワードとしてその 3 つのキーワードが含まれる URL で実験を行った．

判定された順序関係

- {SARS, corona } は 3 つの URL のうち 2 つが共起関係であった．原因としては，SARS が流行してからコロナウィルスが原因であると判明したため，ほぼ同じ時期に出現したと考えられる．
- “ virus ”は医学や公衆衛生のサイトでは長い区間出現していて非依存関係と判定された．

出力結果 図 9 はキーワードの出現状況と出力される時区間を示している．共起関係による質問は，キーワードが出現している近傍の区間も抽出するので全てのキーワードが共起していなくても出力区間となり，従来の AND 条件や OR 条件では取得できない出力を可能とする．

キーワードの頻度とユーザの意図 図 10 は指定した URL(図 9 の 1-c)(注7)のキーワードの出現頻度を示している．3 つのキーワードの頻度が 2003 年の冬から春にかけてピークとなり，夏で 0 になっている．以下の 2 点で図 9 の 1-c の出力結果がユーザが意図する情報であると判定できる．

- 全てのキーワードを含んでいなくても，ほぼそれに近い状態であればユーザが意図する情報を有している可能性は高い
- 図 9 の出力区間とキーワードの頻度が大きい区間が一致している

[実験 1.2] { お中元, お歳暮, ギフト } を質問キーワードとしてその 3 つのキーワードが含まれる URL で実験を行った．

判定された順序関係

- { お中元，お歳暮 } は反復関係であると判定された．日本では夏にお中元，年末の冬にお歳暮を送る習慣があることから，反復関係は妥当であると考えられる．
- 「ギフト」は非依存関係であると判定された．母の日やクリスマスなどのあらゆる時期に出現するためであると考えられる．

出力結果 図 11 はキーワードの出現状況と出力される時区間を示している．2-b を除いた 2 つの URL では「お中元 (midyear)」と「お歳暮 (year-end)」が交互に出現し，それに伴い出力区間も交互になっているのが分かる．異なるハッチングの矩形時区間がクラスタリングされた出力区間を表し，一方が夏の期間，もう一方が冬の期間となっている．その区間に「ギフト」を含んでいるページが出力となるので，出力結果はお中元に相応しいページとお歳暮に相応しいページにクラスタリングされる．

関連キーワードとユーザの意図 我々は，指定した URL(注8)について夏に関するキーワードと冬に関するキーワードの出現状況を調査した．

お中元: 「浴衣」，「水着」，「七夕」,「T シャツ」，「お中元」,「夏休み」,「サマーギフト」,「夏」,「サマーコスメ」,「扇子」

お歳暮: 「温泉」，「お歳暮」，「クリスマス」，「クリスマスツリー」，「バレンタインデー」,「ブーツ」,「行く年」,「福袋」,「元旦」,「ニット」,「コート」,「雪」,「サンタクロース」

図 12 では，時系列ページ各々について夏と冬に関するキーワードの出現数を表している．この場合，キーワード「お中元」は夏の品目に関係があるといえ,「お歳暮は」冬の品目に関係があるといえる．以下の 2 点で図 11 の 2-b の出力結果がユーザの意図する情報であると判定できる．

- 夏の品目に関係があるキーワードと冬の品目に関係があるキーワードが交互に出現している．
- 図 11 の出力区間が交互となっており，関連キーワードの出現数が大きい区間と一致している．

5.2.2 実　　験 2

判定された順序関係から生成される質問によって，複数の URL をクラスタリングする実験を行った．順序関係とクラスタリング結果を表 2 に示す．また，{ 北海道, 沖縄, 旅行 } の質問で，クラスタリングされた 5 つの URL の考察を行った．結果を以下に示す．

- 同じデパートというカテゴリ(注9)の Web ページであるにもかかわらず，クラスタが分かれた．

(注7)：http://www.vdh.state.va.us

(注8)：http://www.hankyu-dept.co.jp/

(注9)：Yahoo!カテゴリに基づく (http://dir.yahoo.co.jp/).

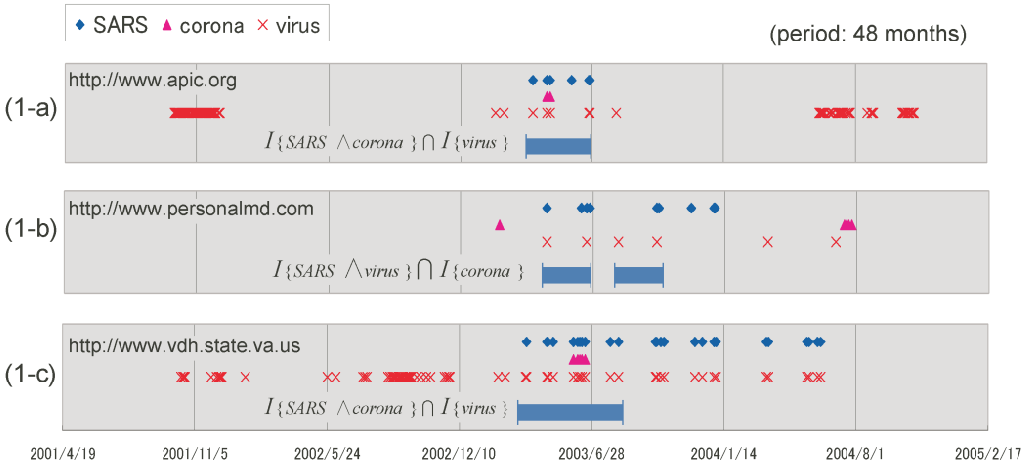

図 9　出力区間　[実験 1.1]

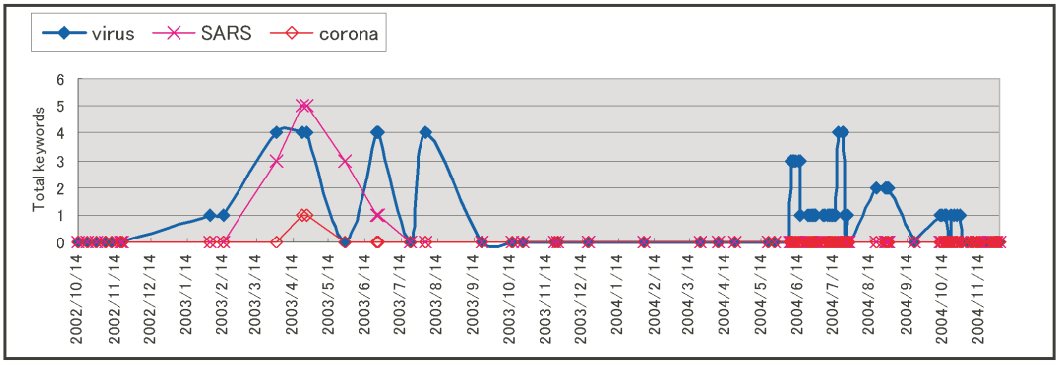

図 10　ページにおける質問キーワードの頻度

- 反対に異なるカテゴリに分類されている Web ページが同じクラスタとなった．

考察を以下に述べる．

- 実際に Web ページを見てみると，旅行会社と同じクラスタとなっているデパートは，北海道から沖縄までホテルや旅館といった友の会優待提携施設があり，テンポラルに旅行プランが掲載されていることが分かった．
- 旅行会社とは別のクラスタとなっているデパートは，関連企業としてトラベル会社はあるが，デパート自体には旅行に関するサービスが存在していなかった．
- 保健研究所も上記同様，旅行に関する情報は掲載されておらず，旅行に関係ないページであった．

同じデパートの Web ページでも，旅行に関係のあるページとそうでないページにクラスタリングされたことが確認できた．したがって，本手法は従来の分類とは異なる「キーワードに基づく動的なクラスタリング」が可能であるといえる．

6. おわりに

我々は，質問キーワードの順序関係を利用し，Web アーカイブの URL と Web ページをクラスタリングして呈示する Web アーカイブの検索手法を提案した．順序関係は，質問キーワードを含むページに基づき時区間を演算し，時系列ページにおける順序関係のある時区間の割合を出すことによって判定した．順序関係によって URL をクラスタリングし，順序関係に基づく質問を再構築することによって Web ページのクラスタリングを行った．クラスタリングの有効性は以下のとおりである．

- 定まった分類手法とは違い，キーワードに基づき，その捉え方が等しい Web ページに分類することができる．
- Web アーカイブに格納されている大規模な Web ページから，ユーザに検索結果を見つけやすく支援することができる．
- ユーザ自身が入力したキーワードを利用するため，ユーザの意図を反映しやすい．

今後の課題は以下のとおりである．

- クラスタリングした Web ページの呈示方法
- データを増やした実験と順序関係判定アルゴリズムの改良
- 局所性や階層性を考慮した，より複雑な順序関係の判定方式の確立
- 同じ Web ページであることの ID 問題の解決

謝　　辞

本研究の一部は，平成 17 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．

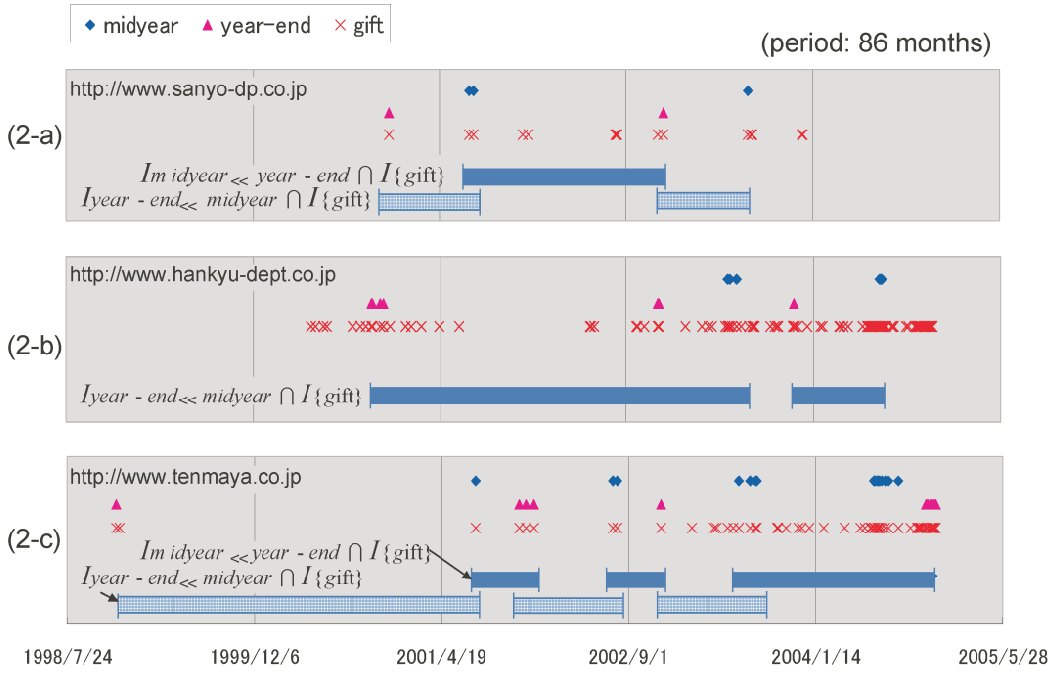

図 11　出力区間　[実験 1.2]

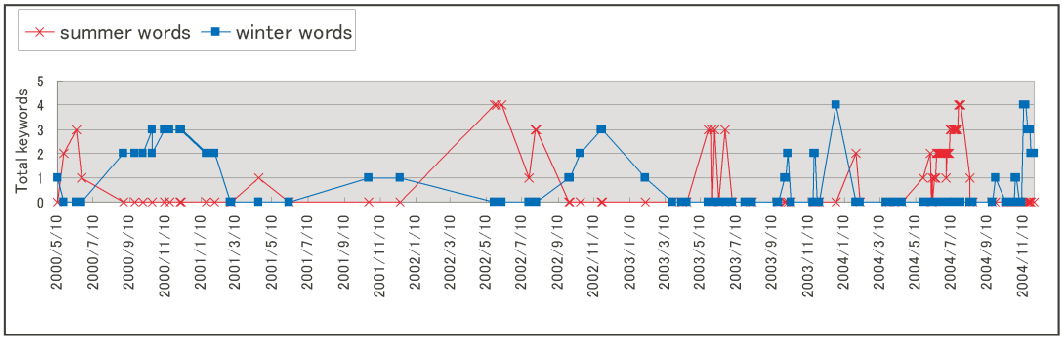

図 12　ページにおける関連キーワードの総数

文　　献

[1] Day, M.: Preserving the fabric of our lives: a survey of Web preservation initiatives, *Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries (ECDL2003)*, pp. 17–22 (2003).

[2] Yan, H., Huang, L., Chen, C. and Xie, Z.: A New Data Storage and Service Model of China Web InfoMall1, *Proceedings of the 4th International Web Archiving Workshop (IWAW04) of 8th European Conference on Research and Advanced Technologies for Digital Libraries* (2004).

[3] Christensen, N. H.: Towards format repositories for web archives, *Proceedings of the 4th International Web Archiving Workshop (IWAW04) of 8th European Conference on Research and Advanced Technologies for Digital Libraries* (2004).

[4] 小城正士，廣瀬信己，河野浩之：Web アーカイブにおける時系列参照アルゴリズムの提案，第 16 回データ工学ワークショップ DEWS'05，電子情報通信学会 (2005).

[5] 豊田正史，喜連川優：日本のウェブアーカイブにおけるコミュニティ発展過程の詳細分析，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).

[6] Toyoda, M. and Kitsuregawa, M.: WebRelievo: A System for Browsing and Analyzing the Evolution of Related Web Pages, *Proceedings of the 3rd International Workshop on Web Dynamics, Part of the workshops track at WWW 2004* (2004).

[7] Otsuka, S., Toyoda, M., Hirai, J. and Kitsuregawa, M.: Extracting User Behavior by Web Communities Technology on Global Web Logs, *Proceeding of 15th International Conference on Database and Expert Systems Applications (DEXA2004)* (2004).

[8] Jatowt Adam, K. K. B. and Ishizuka, M.: Change Summarization in Web Collections, *Proceedings of the 17th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, pp. 653–662 (2004).

[9] Adam, J. and Ishizuka, M.: Web page summarization using dynamic content, *Proceedings of the 13th International World Wide Web Conference (poster session)*, pp. 344–345 (2004).

[10] Adam, J. and Ishizuka, M.: Summarization of Dynamic Content in Web Collections, *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp. 245–254 (2004).

[11] Adam, J. and Ishizuka, M.: Temporal Web Page Summarization, *Proceedings of the 5th International Conference on Web Information Systems Engineering, Brisbane, Australia*, pp. 303–312 (2004).

表 2　判定された順序関係とそれに基づくクラスタリング　[実験 2]

Query Keywords	URL	Relation	
{SARS, corona, virus}	http://www.personalmd.com	非依存, 共起	{SARS, corona}, {corona, virus}, virus ⊕ SARS
	http://www.apic.org	共起,	SARS ⊕ corona, {corona, virus}, {virus, SARS}
	http://www.vdh.state.va.us	非依存	SARS ⊕ corona, {corona, virus}, {virus, SARS}
{ お中元, お歳暮, ギフト }	http://www.hankyu-dept.co.jp	前後，非依存	お歳暮 < お中元, { お歳暮, ギフト }, { ギフト, お中元 }
	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, お歳暮 }
	http://www.tenmaya.co.jp/		お中元 ∘ お歳暮, { お歳暮, ギフト }, { ギフト, お歳暮 }
{ 北海道, 沖縄, 旅行 }	http://www.jtb.co.jp/	反復	北海道 ∘ 沖縄, 沖縄 ∘ 旅行, 旅行 ∘ 北海道
	http://www.iace.co.jp/		北海道 ∘ 沖縄, 沖縄 ∘ 旅行, 旅行 ∘ 北海道
	http://www.mitsukoshi.co.jp/		北海道 ∘ 沖縄, 沖縄 ∘ 旅行, 旅行 ∘ 北海道
	http://www.tenmaya.co.jp/	共起	北海道 ⊗ 沖縄, 沖縄 ⊕ 旅行, 旅行 ⊕ 北海道
	http://www2.pref.shimane.jp/hokanken/		北海道 ⊗ 沖縄, 沖縄 ⊗ 旅行, 旅行 ⊗ 北海道
{ 牛丼 [めし], 豚丼, 定食 }	http://www.zensho.com/	前後，非依存	牛丼 < 豚丼, 定食 < 豚丼, { 定食，牛丼 }
	http://www.yoshinoya-dc.com/		牛丼 < 豚丼, 定食 < 豚丼, { 定食，牛丼 }
	http://www.matsuyafoods.co.jp/		牛めし < 豚丼, 定食 < 豚丼, { 定食，牛めし }
{ 東証，上場, 株式 }	http://www.sundrug.co.jp/index_ f.html	共起,	東証 ⊗ 上場, 上場 < 株式, 東証 < 株式
	http://www.poplar-cvs.co.jp/	前後	東証 ⊕ 上場, 上場 < 株式, 東証 < 株式
	http://www.kirindo.co.jp/	共起，反復，非依存	東証 ⊗ 上場, 上場 ∘ 株式, { 株式, 東証 }
{ 交響曲，ベートーヴェン, 九 }	http://www.city.yokohama.jp/me/mmhall/	共起，反復	交響曲 ⊗ ベートーヴェン, ベートーヴェン ∘ 九, 九 ∘ 交響曲
	http://www.ojihall.com/	前後，共起，反復	交響曲 < ベートーヴェン, ベートーヴェン ⊗ 九, 九 ∘ 交響曲
	http://www.nhk-sc.or.jp/nhk_ hall/	共起	交響曲 ⊗ ベートーベン, 九 ⊗ ベートーベン, 九 ⊗ 交響曲
{ 桜，紅葉，城 }	http://www.jtb.co.jp/	反復	桜 ∘ 紅葉, 紅葉 ∘ 城, 城 ∘ 桜
	http://www.yomiuri-ryokou.co.jp/		桜 ∘ 紅葉, 紅葉 ∘ 城, 城 ∘ 桜
	http://www.orion-tour.co.jp/	反復，前後	桜 ∘ 紅葉, 城 < 紅葉, 城 < 桜
{ 休暇，海外，空港 }	http://www.nishitetsutravel.jp/	前後，反復	海外 < 休暇, 海外 ∘ 空港, 空港 < 休暇
	http://www.yomiuri-ryokou.co.jp/	前後，反復	海外 < 休暇, 海外 ∘ 空港, 空港 ∘ 休暇
	http://www.tobutravel.co.jp/	反復	休暇 ∘ 海外, 海外 ∘ 空港, 空港 ∘ 休暇

[12] 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).

[13] Kleinberg, J.: authorative sources in a hyperlinked environment, *Journal of ACM*, Vol. 48, pp. 604–632 (1999).

[14] Lim, S. J. and Ng, Y.-K.: An Automated Change-Detection Algorithm for HTML Documents Based on Semantic Hierarchies, *Proceedings of the 17th International Conference on Data Engineering*, pp. 303–312 (2001).

[15] 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).

[16] 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).

[17] 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).

[18] 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).