2005－DBS－137（Ｉ）（13）
2005／7／13

Webアーカイブのための 質問キーワードの順序依存を考慮した時系列ページ検索

賀家智代[†]　角谷和俊[††]

現在，Web ページを効率的に収集し Web アーカイブを構築する技術が開発されている．既にいくつかの大規模な Web アーカイブが稼動しているが，Web アーカイブの利用という意味では，単に過去の Web ページを取り出す機能が提供されているのみである．本研究では，Web アーカイブに格納された時系列ページの時間的特性を考慮した検索方式を提案する．本稿では，時系列の順序関係に着目し，ユーザによって入力された複数のキーワードとコンテンツ・トピックの時間的特性から質問意図を抽出する方法について述べる．また，順序関係に基づき，Web アーカイブに対する質問を生成する方法，および，順序依存したページの出力方法について 説明する．さらに，提案する方式に基づく予備実験とプロトタイプシステムの設計について検討する．

A Search Method for Web Archives based on Query Keywords Order

Tomoyo Kage ,[†]　and Kazutoshi Sumiya[††]

Recently, Web pages are effectively collected as Web archives from all over the world, and several methods to construct Web archives have been developed. Although some large-scale Web archives have been operated, they have only the functionalities which users can access the stored Web pages. In this research, we propose a method to retrieve time-series Web pages in Web archives based on temporal characteristics. We focus on time-series order. In this paper, we describe a method to extract the users' intentions using both query keywords and the temporal features of the content topics. We explain a method to generate queries based on the order and a method to output time-series Web pages. Furthermore, we discuss the experiment and the implementation issues of our prototype system based on proposed methods.

1. はじめに

近年，Web アーカイブの構築が盛んに行われ，知的財産として Web ページの収集・保存が行われている．Web アーカイブは，リンク切れや更新などによって失われる情報を永久に保存し，有用な情報を残す機構として注目されている．現在の研究では，Web ページの効率的な収集・保存のための技術や，データベースに格納されて直接アクセスできないデータの取得技術など，特に Web アーカイブ構築手法に主眼が置かれている．

一方，Web アーカイブに蓄積された膨大な時系列データを活用し，そのデータを有効利用しようとする試みも開始されている．大規模データからのデータマイニング[1] や異なる時刻に収集されたコンテンツ間の時間一貫性を保ち，コンテンツの同一性を保障するための方式[2] などが提案されている．しかしながら，Web アーカイブを検索対象として捉え，時系列を考慮して情報を取得する試みは少ない．

Web に対する情報取得については，ページに含まれるキーワードの重要度やリンクの相互参照関係などに基づく効率の良い方式が提案され，検索エンジン技術として既に使用されている．しかしながら，この検索エンジン技術を Web アーカイブに適用しても，検索質問の結果として返されるページ数がアーカイブされた Web の量に比例するだけで，利用価値は少ないと考えられる．

† 兵庫県立大学大学院環境人間学研究科
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

そこで，本研究では Web アーカイブを Web の時系列データと捉え，格納されているコンテンツに含まれる情報 (キーワード) の順序関係に着目し，ユーザの質問意図とコンテンツ・トピックの時間的特性を考慮した検索方式を提案する．本研究では，ユーザが入力した複数の質問キーワードが時間に依存するキーワードかどうかを判定し，順序依存するキーワードと順序非依存のキーワードの組み合わせを自動的に生成する．順序依存するかどうかについては，すべてのキーワードの半順序関係を生成し，それぞれについて順序関係（前後関係や循環関係）であるかどうかについて判定を行う☆．さらに，得られた各々の組み合わせの順序依存性に応じてキーワード質問を自動生成し，順序依存の種類によってキーワードの組み合わせ方法を変えることで，Web アーカイブに含まれる時系列ページから，ユーザが意図するページを取得する．

本稿では，順序依存を考慮した時系列ページ検索のための質問キーワードの順序依存判定方式，および順序関係に基づく質問生成について議論を行う．以下の構成は次の通りである．2 節では，従来のアプローチと本研究の概要について述べる．3 節では質問キーワードの順序依存判定の手順について説明し，4 節では得られた順序関係に基づいた質問生成法について述べる．さらに，5 節では提案する方式に基づいた予備実験とプロトタイプシステムの設計について述べ，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 従来のアプローチ

Web アーカイブに格納された Web ページは，現在の Web では得られない情報が取得可能である．例えば，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[3] や，ある時間に依存した情報など，様々な観点での検索が可能となる．しかしながら，従来の検索方式では時区間に依存する検索は難しい．

例えば，Web ページを時間順に並べたリストを P，キーワード k を含むページの集合を P_k，入力キーワードの集合を K，得られる解を A とする．

$P = p_1\ p_2 \cdots p_{10}$

$P_{\text{サマーバーゲン}} = \{p_1, p_7\}$

$P_{\text{ウィンターバーゲン}} = \{p_4, p_{10}\}$

$K = \{\text{ウィンターバーゲン}, \text{サマーバーゲン}\}$

通常，キーワード質問による Web ページの検索は，

☆ キーワードの組み合わせによる順序依存性であるため，同一キーワードであっても，その組み合わせにより他方のキーワードが変わった場合は依存性の有無が変更される場合がある

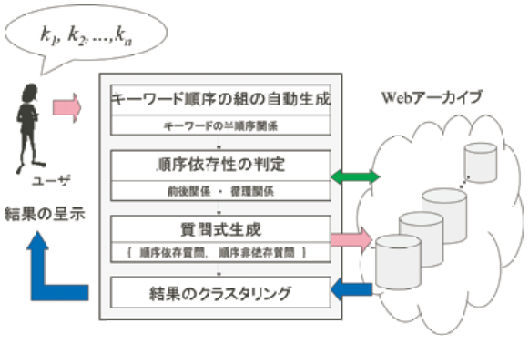

図 1 本研究の概要

そのキーワードを含む Web ページが解となる．また，複数のキーワードによる検索では，AND 条件や OR 条件を用いる．この例では，AND 条件で質問を生成すると，$Q =$ (サマーバーゲン $\wedge$ ウィンターバーゲン) となり，得られる解はない．一方，OR 条件で質問を生成すると，$Q =$ (サマーバーゲン $\vee$ ウィンターバーゲン) となり，解は $A = \{p_1, p_4, p_7, p_{10}\}$ となる．

AND 条件による検索は，同じ時刻にキーワードが共起していなければ解とならない．したがって，複数のキーワードが異なる時間に存在するような場合は解が得られない．そのため，時系列ページのような時区間を持ったデータを検索する場合は，AND 条件での問い合わせだけでは不十分である．

OR 条件による解は $\{p_1, p_4, p_7, p_{10}\}$ となり，キーワード単独の情報かつ断片的である．異なるキーワードのページを断片的に取得しただけでは，解が何を表しているか分かりにくく，質問の意図が十分に反映した解とはならない場合が多い．

2.2 本研究の概要

本研究は，同一 URL の異なる時間の Web コンテンツ☆☆，すなわち同一 URL の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☆☆☆．概要を図 1 に示し，以下に述べる．

まず，ユーザが入力した複数の質問キーワードが時間に依存するか否かを判定するために，2 つのキーワードの組を自動的に生成する．すべての組について時系列ページにおけるキーワード出現順序を調べる．この時，順序については出現時間の半順序関係 (partial

☆☆ ブックマークや Wayback Machine[4] の出力結果など．

☆☆☆ 本稿では，同一 URL に限定して議論を行うが，厳密に同一の URL を持つページでなくとも同じトピックに関する時系列ページであれば，同様の処理は可能であると考えられる．

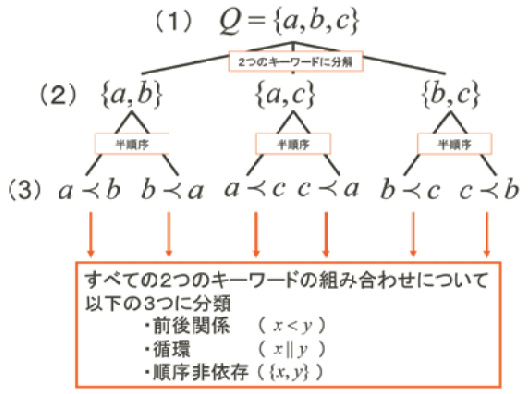

図 2 質問キーワードの順序依存判定の手順

order) に基づき，条件式を生成する．

次に，順序依存性の判定のために，半順序関係が保たれている区間および半順序関係に矛盾が生じる区間をそれぞれ計算する．これらの区間を比較することにより，2 つのキーワードに順序依存性があるかどうかを判定する（このとき，あるキーワードは他のキーワードよりも前に出現するという前後関係，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循環関係，あるいは，順序依存性がないという順序非依存関係の 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，順序依存性が変化することに注意を要する．

さらに，得られた各々の組み合わせの順序依存性に応じて質問式を自動生成する．この時，順序依存の種類によってキーワードの組み合わせ方法を変えることで，Web アーカイブに含まれる時系列ページから，ユーザが意図するページを取得することが可能となる．

このように，本方式はキーワードの出現順序の半順序性に着目する．出現順序の半順序性とは，ある程度の順序が決まっていても順序が一意に決まらず，これよりは後，これよりは前といった順序の性質のことをいう．時間に依存して出現するキーワード，例えば季節やイベント，行事を表すキーワードは，大まかに順序が決まっている．しかし，絶対的な表現とは異なり，キーワードがページに出現する時間にはズレが生じる．そのため順序が一意に決まらず，出現順序は半順序の性質を持つといえる．

関連研究としては，Web ページのマルチメディア検索における質問緩和[5),6)] やビデオの時区間抽出のグルー結合[7)] などがあるが，本研究では時系列 Web ページを対象としているため，これらの方式とは検索対象が異なる．また，時系列 Web ページのための検索エンジンとして blogWatcher[8),9)] があるが，対象

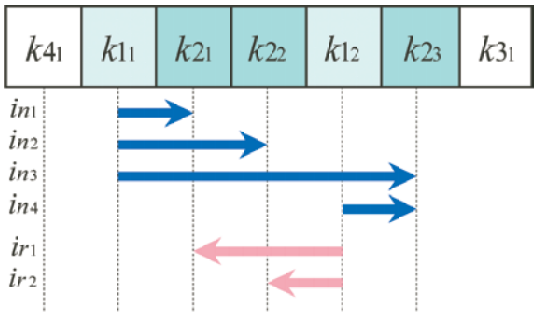

図 3 正順序区間と逆順序区間の抽出

を blog に特化している点で本研究と異なる．

3. 質問キーワードの順序依存

3.1 順序関係によるキーワードの時区間抽出

ユーザが入力した複数のキーワードが，時間に依存するキーワードかどうかを判定するための手順を図 2 に示す．この例の場合，(1) で指定された 3 つのキーワードを (2) のように，2 つのキーワードの組み合わせに分解する．次に，(3) のように 2 つのキーワードが半順序関係となる 2 つの場合に展開する．展開されたすべての半順序関係について，前後関係，循環関係，順序非依存の 3 つの関係に分類する．得られた結果を基に，キーワードの組み合わせ方法を変えて質問を生成する．本研究で対象とするのは，Web アーカイブに格納されている同一 URL の時系列ページである．

任意の 2 つのキーワード k_a, k_b を入力とすると半順序による条件は以下の 2 通りである．

$$k_a \prec k_b, \quad k_b \prec k_a$$

この 2 つの式に対してそれぞれ，式が条件を満たす区間と，条件を満たさない区間を抽出する．直感的に言うと，条件を満たさない区間以外のところが候補区間となる．ただし，この候補区間以外でも条件を満たす区間との重なりや包含関係がある場合は，これを候補区間とする．条件を満たす区間を**正順序区間**，条件を満たさない区間を**逆順序区間**と呼ぶ．

3.2 正順序区間と逆順序区間

3.2.1 正順序区間

正順序区間とは，条件に適合した区間で $\prec$ で結合された 2 つのキーワードのうち**前者**のキーワードを含むページを始端，**後者**のキーワードを含むページを終端とする区間をいう．条件 $k_a \prec k_b$ の正順序区間 i_n は，キーワード k を含むページを $p(k)$ で書き表すと，

$$i_n = [p(k_a), p(k_b)]$$

となる．通常，同一キーワードは複数出現するので，正順序区間は 1 つ以上抽出される．i_n の集合を $I_n (I_n =$

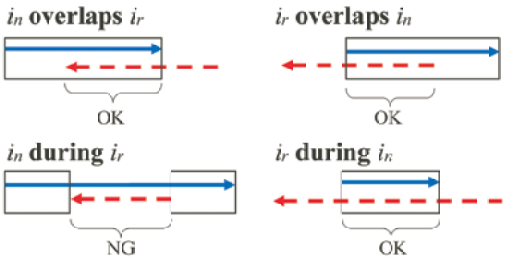

図 4　正順序区間と負順序区間の関係

$\{i_{n_1}, i_{n_2}, ..., i_{n_n}\}$) とする．図 3 では，$k_1$ から k_4 を含む時系列 Web ページに対する条件 $k_1 \prec k_2$ の正順序区間を右矢印で示している．この場合，正順序区間は以下の 4 つの区間となる．

$I_n = \{i_{n_1}, i_{n_2}, i_{n_3}, i_{n_4}\}$
$i_{n_1} = [p(k_{11}), p(k_{21})]$, $i_{n_2} = [p(k_{11}), p(k_{22})]$
$i_{n_3} = [p(k_{11}), p(k_{23})]$, $i_{n_4} = [p(k_{12}), p(k_{23})]$

ここで，k_{i_j} は時系列ページの中で，キーワード k_i が j 番目に出現したことを表している．

3.2.2　逆順序区間

逆順序区間とは，条件に適合しない矛盾した区間で，$\prec$ で結合された 2 つのキーワードのうち**後者**のキーワードを含むページを始端，**前者**のキーワードを含むページを終端とする区間をいう．条件 $k_a \prec k_b$ の逆順序区間 i_r は，

$i_r = [p(k_b), p(k_a)]$

逆順序区間も正順序区間と同様に 1 つ以上抽出される．i_r の集合を $I_r (I_r = \{i_{r_1}, i_{r_2}, ..., i_{r_m}\})$ とする．

図 3 では，逆順序区間を左矢印ので示している．この場合，逆順序区間は以下の 2 つとなる．

$I_r = \{i_{r_1}, i_{r_2}\}$
$i_{r_1} = [p(k_{21}), p(k_{12})]$, $i_{r_2} = [p(k_{22}), p(k_{12})]$

3.3　時区間抽出

3.2 節の条件に基づく 2 つの区間から，条件 $k_a \prec k_b$ のページの時区間抽出を行う．時区間抽出の手順は以下の通りである．

1. 逆順序区間以外の時区間を解の候補とする．
2. ただし，逆順序区間に含まれる区間でも，正順序と見なされる区間は解とみなす．

順序条件によって抽出する時区間は，その条件に反しない区間，すなわち逆順序区間を検索対象から取り除いた区間となる．しかしながら，それだけでは検索対象の大部分または全てを逆順序区間が占めている場合に，意図した解が得られない．そこで，逆順序区間であっても条件に適合する区間も解とみなし抽出する．

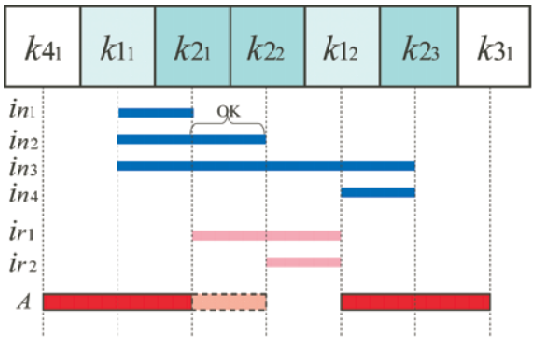

図 5　時区間抽出の例

3.4　時区間抽出の手順

まず，検索対象となる時系列ページ P から逆順序区間以外を抽出する．例えば，図 3 及び図 5 では，時系列 Web ページにおける逆順序区間以外の区間は I_r を含まない区間，すなわち図 5 の A の実線区間の $[p(k_{41}), p(k_{21})]$ と $[p(k_{12}), p(k_{31})]$ が該当する．

図 4 は，正順序区間 i_n と逆順序区間 i_r の関係を J.F.Allen の区間論理[10] `overlaps`（上段）と `during`（下段）によって表したものである．4 つの図はそれぞれ i_n を実線，i_r を点線，質問の解を四角枠で表現している．この 4 つの区間関係

i_n `overlaps` i_r (1)
i_r `overlaps` i_n (2)
i_n `during` i_r (3)
i_r `during` i_n (4)

のうち，逆順序区間であっても解とする場合がある．すなわち，式 (1), (2), (4) の場合である．注目すべき点は，`overlaps` は i_n と i_r を入れ替えても成立するが，`during` は成立しないことである．式 (4) は，通常解とはならない区間に質問に適した区間が含まれるため，その区間は解となる．しかし，式 (3) は，質問に適した区間に矛盾する区間が含まれるため，解とはならない．なお，これらの関係 (式 (1), (2), (4)) が複数出現した場合は，`overlaps` より `during` を，`during` の中でも区間が短いものをそれぞれ優先して解とする．

図 5 では，逆順序区間と正順序区間が重なる関係が i_{n_2} `overlaps` i_{r_1}, i_{n_3} `during` i_{r_1}, i_{n_3} `during` i_{r_2} の 3 つ存在する．上記のうち，解となる関係式 (1), (2), (4) に該当するのは，i_{n_2} `overlaps` i_{r_1} であるため，$[p(k_{21}), p(k_{22})]$ となる（A の点線区間）．したがって，図 3 における $k_1 \prec k_2$ の時区間は，図 5 に示すように，$[p(k_{41}), p(k_{22})]$ と $[p(k_{12}), p(k_{31})]$ の 2 つの区間である．

3.5 キーワードが共起する場合

順序関係のあるキーワードが 1 つのページに共起する場合がある．本手法では，順序関係があるキーワードが同じ時点に出現しても，順序には反しないとして同様に扱う．また，`overlaps` と `during` の他に `starts`, `finishes`, `equals` も考慮し，逆順序区間の一部が正順序区間である場合を解とする．ただし，`equals` は正順序でもあり逆順序でもある区間なので解としない．

4. 順序関係に基づく質問生成

4.1 順序依存の種類

キーワードの順序依存とは，2 つのキーワードの順序関係が前後の一方または両方に決定される関係をいう．キーワード数を n とすると生成される式は ${}_nP_2$ 通りである．生成されたすべての式に対して，順序依存の種類を判定する．基本的に以下の 3 種類に分類される．

1. 前後関係　：順序が一意に決まる場合
2. 循環　　：交互に現れる場合
3. 順序非依存：順序関係がない場合

4.2 順序依存判定

4.2.1 前後関係の判定

まず，キーワード k_a, k_b の半順序に基づく条件式 $k_a \prec k_b$, $k_b \prec k_a$ をそれぞれ問い合わせる．そして両者の解の数を比較し，閾値（大きい値）以上の偏りがある場合，解の数が多い条件式の方が k_a, k_b の順序を適切に表現していると考えられるため，その式を質問の要素とする．

4.2.2 循環または順序非依存の判定

前後関係がないと判断された場合，k_a, k_b は順序非依存か，循環しているかのどちらかである考えられる．

循環している場合は，$k_a \prec k_b$ と $k_b \prec k_a$ で得られる時区間が互いに排他的になると考えられる．したがって，両者の AND 条件で得られる時区間が閾値（小さい値）よりも小さい場合，k_a と k_b は循環していると判断し，2 つの条件式は共に質問式の要素となる．順序依存しない場合は，$k_a \prec k_b$ と $k_b \prec k_a$ で得られる時区間の共通部分が多くなると考えられる．したがって，両者の AND 条件で得られる時区間が閾値（小さい値）よりも大きい場合，k_a と k_b は順序に関係がなく，順序非依存の関係である．

$k_a \prec k_b$ の解の集合を $I_{k_a \prec k_b}$ とし，$I_{k_a \prec k_b} \cap I_{k_b \prec k_a}$ が検索対象の全区間に占める割合を**網羅度**と呼ぶ．網羅度とは，検索対象区間に占める区間の割合であり，どのくらい区間を網羅しているかを表す．4 節で述べた以上の手順を示す．

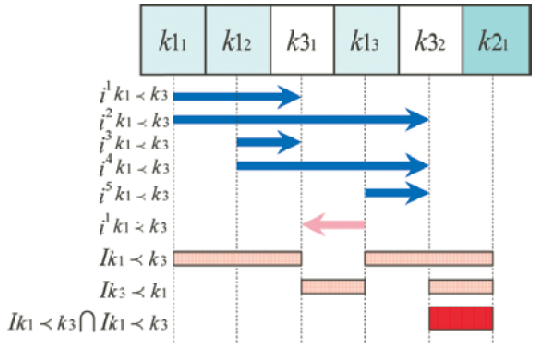

図 6　循環性の抽出

1. **$k_a \prec k_b$, $k_b \prec k_a$ をそれぞれ問い合わせる．**
2. **$k_a \prec k_b$ の解の数と $k_b \prec k_a$ の解の数を比較する．**
3. **解の数に偏りがある場合は多い方の式を質問の要素とする（$Q_{k_a<k_b}$ または $Q_{k_b<k_a}$）．解の数に偏りがない場合は次の処理に進む．**
4. **$I_{k_a \cap k_b}$ AND $I_{k_b \cap k_a}$ の網羅度を計算し，値が小さい場合は両者とも質問の要素となる（$Q_{k_a \| k_b}$）．値が大きい場合は 2 つのキーワードは順序非依存の質問となる（$Q_{\{k_a,k_b\}}$）．**

ここで，$Q_{k_a<k_b}$ または $Q_{k_b<k_a}$ は k_a, k_b の前後関係を表す質問の要素，$Q_{k_a \| k_b}$ は k_a, k_b の循環関係を表す質問の要素，$Q_{\{k_a,k_b\}}$ は k_a, k_b が順序非依存関係を表す質問の要素である．

図 6 は，$k_a \prec k_b$ の正順序区間を $i_{k_a \prec k_b}$（右矢印），逆順序区間を $i_{k_a \succ k_b}$（左矢印）で表現している．時区間 $I_{k_a \prec k_b}$ は，$[p(k_{11}), p(k_{31})]$, $[p(k_{13}), p(k_{21})]$，$I_{k_b \prec k_a}$ は，$[p(k_{31}), p(k_{13})]$, $[p(k_{32}), p(k_{21})]$ であるため，☆．$I_{k_1 \prec k_3 \cap k_3 \prec k_1}$ の区間は 1 となり，網羅度は小さい．したがってこれらは循環関係にあるといえる．

4.3 質 問 生 成

ユーザが入力したキーワードをすべて含むように条件式の組み合わせを行い質問を生成する．前後関係を表す式，循環する式，順序非依存式は質問の要素となる条件式である．

- 前後関係を表す質問　：$Q_{k_a<k_b}$ または $Q_{k_b<k_a}$
- 循環を表す質問　　：$Q_{k_a \| k_b}$
- 順序非依存を表す質問：$Q_{\{k_a,k_b\}}$

前節の例では，$k_1 \prec k_2$ と $k_3 \prec k_2$ が前後関係，k_1 と k_3 が循環を表すので，$Q_{k_1<k_2}, Q_{k_3<k_2}, Q_{k_1 \| k_3}$ の組み合わせとなり，質問は，2 つの要素を組み合わせ

☆ $k_a \prec k_b$ の正順序区間は $i_{k_a \prec k_b}$，逆順序区間は $i_{k_a \succ k_b}$ と考えられるため

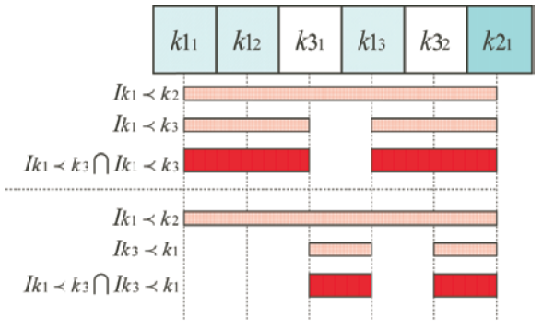

図 7 $Q_{k_1<k_2} \ \wedge \ Q_{k_1 \| k_3}$ の解

る以下の 3 通りである．

$$Q_{k_1<k_2} \ \wedge \ Q_{k_3<k_2}$$
$$Q_{k_1<k_2} \ \wedge \ Q_{k_1 \| k_3}$$
$$Q_{k_3<k_2} \ \wedge \ Q_{k_1 \| k_3}$$

この場合，2 番目の $Q_{k_1<k_2} \ \wedge \ Q_{k_1 \| k_3}$ の解は以下の通りとなる（図 7 参照）．

$$\begin{aligned} A(Q_{k_1<k_2} \ \wedge \ Q_{k_1 \| k_3}) &= I_{k_1 \prec k_2} \cap (I_{k_1 \prec k_3} \cup I_{k_3 \prec k_1}) \\ &= (I_{k_1 \prec k_2} \cap I_{k_1 \prec k_3}) \cup (I_{k_1 \prec k_2} \cap I_{k_3 \prec k_1}) \end{aligned}$$

図 7 は，$I_{k_1 \prec k_2}$ と図 6 の $I_{k_1 \prec k_3}$，$I_{k_3 \prec k_1}$ をそれぞれ AND で結合した結果で，解を表している．上部の図が展開式の $\cup$ で結合されている式の前者 $(I_{k_1 \prec k_2} \cap I_{k_1 \prec k_3})$，下部の図が後者 $(I_{k_1 \prec k_2} \cap I_{k_3 \prec k_1})$ で，解の一部を表している．

すべてのキーワードの組が順序依存する場合，出現する順序を適切に表した質問を問い合わせることが望ましい．問い合わせた解のうち，より長い時区間を占めている質問がキーワードの順序を適切に表せていることになる．そのため，より的を絞った解を得るために，網羅度が極端に大きい場合は更に条件式を組み合わせて質問を再構築する．

5. 予備実験とプロトタイプシステム

5.1 予 備 実 験

5.1.1 実　　験

本方式を評価するために，予備実験を行った．あるショッピングモールのページ（http://the.mall-himeji.jp/），2002 年 5 月から 2004 年 6 月までの 29 ページを用いて実験を行った．単語及び時区間の抽出を手動で行い，ページに含まれるテキストが少ないことから画像に含まれる単語も考慮した．質問キーワードは「サマーバーゲン」,「クリスマスフェア」で問い合わせた．

5.1.2 結果と考察

$Q =$ { サマーバーゲン, クリスマスフェア | サマーバーゲン $\prec$ クリスマスフェア } の結果は，2002 年 5 月から 12 月，2004 年 8 月から 2005 年の 6 月の 2 つの区間のページが解となった．$Q =$ { サマーバーゲン, クリスマスフェア | クリスマスフェア $\prec$ サマーバーゲン } の結果は，2002 年 5 月から 8 月，2002 年 12 月から 2003 年 8 月，2003 年 12 月から 2004 年 6 月の 3 つの区間のページが解となった．

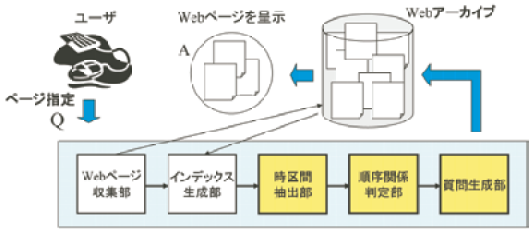

図 8 システム構成図

取得した時区間は，全体の時区間の両端が重複し，中央が排他的な結果となった．原因は，循環している事象にもかかわらず，検索対象が有限であるため最初と最後の区間に逆順序区間が成立しないからだと考えられる．

重複した区間，すなわち解が共通している区間は無関係であると考えられるため，排他的な区間で得られたページの内容の考察を行った．排他的な時区間は，前者では 2002 年 8 月から 12 月，2003 年 8 月から 12 月，後者では 2002 年 12 月から 2004 年 8 月の区間であった．

前者は夏と冬の間で,「秋のショッピングラリー」や「手袋」といった単語が特徴的であった．内容は冬の傾向が強く，クリスマスフェアで売られるものが想定できた．後者は冬と夏の間で,「サマープレゼントフェア」や「サングラス」といった単語が特徴的であった．内容は夏の傾向が強く，サマーバーゲンで売られるものが想定できた．実験の結果，キーワードの順序によって異なる情報が取得できることが確認された．

5.2 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．プロトタイプは現在構築中である．

1. Web **ページ収集部** まず，ユーザによって指定されたページの時系列ページを取得する．取得方法は，InternetArchive☆4) から，そこに張られたリンクによって抽出する．
2. **インデックス生成部** ページのテキストを形態

☆ http://web.archive.org/web/*/任意の URL

素解析し，名詞を抽出する．その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．

3. **時区間抽出部** インデックスを参照し，キーワードによる条件式を基にページの時区間を抽出する．
4. **順序関係判定部** 条件式を生成している 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．

ページ内に含まれるキーワード全てを対象とするには大量の処理が必要であるため，時区間指定のための着目する（インデックス化する）キーワードの範囲を定める必要がある．時系列 Web ページの集合を基に df 値を算出し，*tf-idf* 法[11] によりキーワードの重み付けを行う．キーワードの *tf-idf* 値によってランク付けを行い，範囲を求める．この処理はインデックス生成部にて行う．

6. まとめと今後の課題

本稿では，Web アーカイブに格納された時系列 Web ページの検索において，キーワードの順序に基づく問い合わせ方式を提案した．入力されたキーワードから半順序に基づく条件式を生成し，時区間を抽出することによって，時間的に意味付けされたまとまりのあるページの取得を可能とした．

今後の課題としては，実験のデータを増やしてアルゴリズムを改良し，検索対象の拡張を目的とした，更に実用的な検索方式を検討していく方針である．

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．

参 考 文 献

1) 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 (2003).
2) 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 (2005).
3) Cyclone, *http://cyclone.slis.tsukuba.ac.jp/*.
4) Internet Archive : Way Back Machine, *http://www.archive.org/*.
5) 桑原昭裕, 角谷和俊, 田中克己: 質問緩和法によるクロスメディア・メタサーチ, 第 15 回データ工学ワークショップ DEWS'04 (2004).
6) Kuwabara, A. and Tanaka, K.: RelaxImage:A Cross–Media Meta–Search Engine for Searching Images from Web Based on Query Relaxation, *21st International Conference on Data Engineering(ICDE2005)*, pp.1102–1103 (2005).
7) プラダンスジット, 田島敬史, 田中克己: ビデオデータ検索のための区間グルー操作と解のフィルタリング, 情報処理学会論文誌, Vol. 40, No. SIG3(TOD1), pp. 80–90 (1999).
8) BlogWatcher, *http://blogwatcher.pi.titech.ac.jp/*.
9) 奥村学, 南野朋之, 藤木稔明, 鈴木泰裕: blog ページの自動収集と監視に基づくテキストマイニング, 人工知能学会セマンティックウェブとオントロジー研究会 (SIG-SWO-A401-01) (2004).
10) Allen, J. F.: Maintaining Knowledge about Temporal Intervals, *Communications of the ACM*, Vol. 26, No. 11, pp. 832–843 (1983).
11) Salton, G. and Yang, C. S.: On the specification of tern values in automatic indexing, *J. Documentation*, Vol. 29, No. 4, pp. 351–372 (1973).