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

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

賀家 智代♥　角谷 和俊♠
Tomoyo KAGE　Kazutoshi SUMIYA

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，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．本稿では，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 Web アーカイブの時系列特性を考慮した検索方式を提案する．また，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．

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 retrieve useful information for users. In this paper, we propose a method to retrieve time-series of Web pages from Web archives by analyzing appearance tendencies of keywords on the time-series pages. We have also considered experiment and implementation issues regarding the prototype system.

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]など多くの構築手法が検討されている．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いない．

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」は「あるWeb ページにおけるキーワードの役割」と考え，キーワードの捉われ方の等しい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の抽出を支援する方式を提案する．

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

以下，本稿の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やOR条件が用いられる．しかしながら，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，OR条件での出力は多くの不要なページが含まれ，AND条件では出力が得られない場合がある．

そこで我々は，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 Web ページにおける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．本研究の概要を図1 に示す．

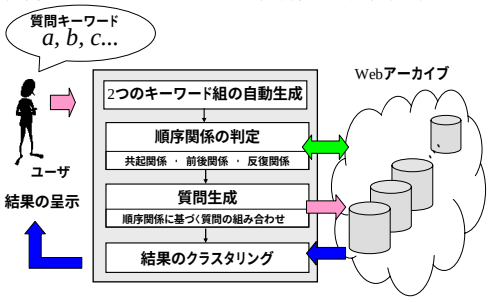

図 1　本研究の概要

Fig.1　Basic Concept

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．順序関係は，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変化する場合がある．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

♥ 学生会員 兵庫県立大学大学院環境人間学研究科博士前期課程 nd05w005@stshse.u-hyogo.ac.jp
♠ 正会員 兵庫県立大学環境人間学部 sumiya@shse.u-hyogo.ac.jp

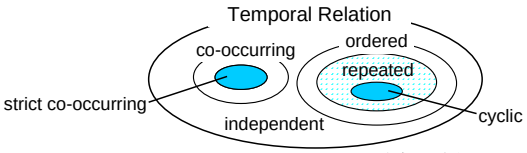

図 2　キーワードの順序関係
Fig.2　Temporal Relation of Keywords

よって，Web ページをクラスタリングし検索結果とする．

2.2 関連研究

Webアーカイブの検索エンジンとして，WayBack Machine[2]やWERA[3]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困難である．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．

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[4]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[5]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．

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[6]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．

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 に示す．なお，順序依存関係を「共起」，「前後」，「反復」とし，非依存関係を合わせて 4 つの基本的順序関係とする．基本的順序関係に「共起」の強い関係である「密共起」，「反復」の強い関係である「循環」を追加して 6 つの順序関係を定義する．

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）時系列ページにおいて，キーワード a, b の一方が他方より先に出現する関係を前後関係という．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」はこの関係である．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）時系列ページにおいて，キーワード a, b が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．反復関係は複数の前後関係（「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）が成立するため前後関係に包含される．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．

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，「スキー」と「スノーボード」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．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．

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．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順序関係を判定する．判定する任意のキーワードを a, b として，図 3 に示し以下に述べる．

3.2.1　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，$I_{a \prec b}$ 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．

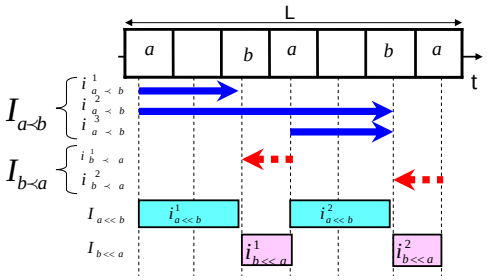

図 3　キーワードの順序関係
Fig.3　Keywords Relation

3.2.2　順序関係判定方法

共起（co-occurring）同時期といえる時区間の閾値を定め，キーワード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．時系列ページに出現するキーワードの総数に占める共起総数の割合が閾値 α より大きければ共起関係

とみなす． 共起関係である場合，共起の範囲を 0（つまり同一ページ）として共起の判定同様に密共起関係を判定する．

前後（ordered）順序関係が保たれている区間（$I_{a \ll b}$ と $I_{b \ll a}$）が時系列ページの時区間(L)に占める割合を算出し，キーワードが順序依存しているか否かを判定する．この値が閾値 β より大きく，$I_{a \ll b}$ と $I_{b \ll a}$ の偏りが閾値 γ 以上であれば前後関係とみなす．

反復（repeated）前後関係の判定同様にキーワードが順序依存しているか否かを判定する．順序依存していれば反復区間として $i_{a \ll b}$ と $i_{b \ll a}$ が交互に出現している区間を算出する． $i_{a \ll b}$ と $i_{b \ll a}$ の総和に占める反復区間の割合が閾値 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は $i_{a \ll b}$ と $i_{b \ll a}$ の各時区間の分散値を計算し，各値が共に閾値 δ よりも小さい場合を循環関係とみなす．

非依存（independent）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．

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．$I_{\{a \vee b\}}$ 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． $Q_{a \circledcirc b}$ は循環関係で，平均周期の時区間を返す． $Q_{a \otimes b}$ と $Q_{\{a,b\}}$ はキーワードの出現状況について問い合わせ， $Q_{a \prec b}$ 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を生成する．

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，c は独立となるため，全てのキーワードを含むように組み合わせた質問は $Q_{a \oplus b} \wedge Q_{\{c\}}$ となり，抽出される区間は以下のようになる．

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

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 $\{c, a\}$ が非依存の場合，質問の要素は $Q_{a \circ b}$ ，$Q_{\{b,c\}}$ 及び $Q_{\{c,a\}}$ となり，全てのキーワードを含むように組み合わせた質問は $Q_{a \circ b} \wedge Q_{\{c\}}$ となり，抽出される区間は以下のようになる．

$$R(Q_{a \circ b} \wedge Q_{\{c\}}) = (I_{a \ll b} \bigcup I_{b \ll a}) \bigcap I_{\{c\}}$$
$$= (I_{a \ll b} \bigcap I_{\{c\}}) \bigcup (I_{b \ll a} \bigcap I_{\{c\}})$$

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．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．

表 1 生成される質問と抽出区間

Table 1 The query elements and the extracted intervals

関係	質問	抽出される区間
共起	$Q_{a \oplus b}$	$I_{\{a \wedge b\}}$
密共起	$Q_{a \otimes b}$	$I_{\{a \wedge b\}}$，同じページ上にキーワード a と b が出現
前後	$Q_{a \prec b}$	$I_{\{a \ll b\}}$
反復	$Q_{a \circ b}$	$I_{\{a \ll b\}} \bigcup I_{\{b \ll a\}}$
循環	$Q_{a \circledcirc b}$	$I_{\{a \ll b\}} \bigcup I_{\{b \ll a\}}$ 各区間の長さは等しい
非依存	$Q_{\{a,b\}}$	$I_{\{a \vee b\}}$

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 4 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．

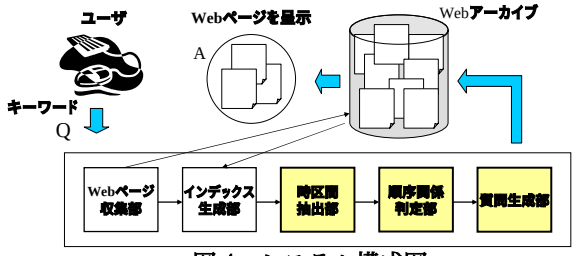

図 4 システム構成図
Fig.4 System Architecture

(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，そのページに対応する時系列の Web ページを InternetArchive(http://web.archive.org/web/*/(任意の URL）) から収集する．ただし，本プロトタイプではアドレスの変化を考慮して 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 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，Internet-Archive(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．

(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 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 の時系列 Web ページを自動的に取得した．各々の質問キーワードを複数の任意の URL に対して問い合わせた．閾値は，予備実験によりそれぞれ α，β，γ を 0.3, 0.6,5，共起の範囲を 90 日とした．

[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．一部の結果を表 2 に示す． {SARS, corona, virus}を質

表 2　生成される質問と抽出区間
Table 2　The query elements and the extracted intervals

質問キーワード	URL	順序関係	
{SARS, corona, virus}	http://www.personalmd.com	非依存, 共起	{SARS, corona}, {corona, virus}, virus⊕SARS
	http://www.apic.org http://www.vdh.state.va.us	共起, 非依存	SARS⊕corona, {corona, virus}, {virus, SARS} 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 http://www.tenmaya.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} お中元○お歳暮, {お歳暮, ギフト}, {ギフト, お歳暮}
{牛丼[めし], 豚丼, 定食}	http://www.zensho.com/ http://www.yoshinoya-dc.com/ http://www.matsuyafoods.co.jp/	前後, 非依存	牛丼 < 豚丼, 定食 < 豚丼, {定食, 牛丼} 牛丼 < 豚丼, 定食 < 豚丼, {定食, 牛丼} 牛めし < 豚丼, 定食 < 豚丼, {定食, 牛めし}
{北海道, 沖縄, 旅行}	http://www.jtb.co.jp/ http://www.iace.co.jp/ http://www.mitsukoshi.co.jp/	反復	北海道○沖縄, 沖縄○旅行, 旅行○北海道 北海道○沖縄, 沖縄○旅行, 旅行○北海道 北海道○沖縄, 沖縄○旅行, 旅行○北海道
	http://www.tenmaya.co.jp/ http://www2.pref.shimane.jp/hokanken/	共起	北海道⊕沖縄, 沖縄⊕旅行, 旅行⊕北海道 北海道⊕沖縄, 沖縄⊕旅行, 旅行⊕北海道

問キーワードとした場合, 順序関係は{SARS, corona}が共起, {virus }が非依存となった. 出力は 2003 年の 3 月から 6 月の 6 ページとなった. 図 5 はある URL におけるキーワードの出現頻度を示している. {SARS, corona}の出現傾向が等しく{virus}のみ異なる傾向であることから, 判定された順序関係は妥当であるといえる. また, 以下の点で出力が有効であることが分かる.

・キーワードの出現頻度が高い区間とシステムが出力したページの区間が合致している.
・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.

[実験 2]　{北海道, 沖縄, 旅行}という質問キーワードから判定された順序関係によって, 5 つの URL をクラスタリングする実験を行った（表 2）. 結果を以下に示す.

•デパートという同じカテゴリの URL であるにもかかわらず, クラスタが分かれた.
• 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.

旅行会社と同じクラスタのデパートの URL 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, 頻度に旅行プランが掲載されていた. 旅行会社とは別のクラスタのデパートは, 旅行に関するサービスが存在しなかった. デパートという同一カテゴリの URL でも旅行に関係のある URL と関係のないものにクラスタリングされた. 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」が可能であると考えられる.

6. おわりに

本稿では, 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.

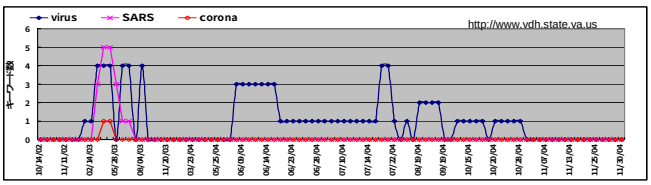

図 5　ページにおけるキーワードの出現頻度
Fig.5　Keywords Frequencies on Web Pages

今後の課題は, 呈示方法, 順序関係判定アルゴリズムの改良や Web ページの ID 問題の解決などが挙げられる.

[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.

[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] Internet Archive：http://www.archive.org/
[3] WERA：http://archive-access.sourceforge.net/projects-/wera/
[4] 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).
[5] 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).
[6] 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).

賀家 智代　Tomoyo KAGE
兵庫県立大学大学院環境人間学研究科博士前期課程在学中. 2005 年姫路工業大学環境人間学部環境人間学科卒業. 情報処理学会, 日本データベース学会学生会員.

角谷 和俊　Kazutoshi SUMIYA
兵庫県立大学環境人間学部環境人間学科教授. 1998 年神戸大学大学院自然科学研究科博士後期課程修了, 博士(工学). マルチメディアデータベース, データ放送の研究開発に従事. IEEE Computer Society, ACM, 電子情報通信学会, 情報処理学会, 日本データベース学会等各会員.