Web アーカイブを利用したブックマークの意味的補間方式

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:　†na01v072@stshse.u-hyogo.ac.jp,　‡sumiya@shse.u-hyogo.ac.jp

あらまし　従来の Web ブックマークはページの URL を保持するのみであるので，以前にブックマークしていたページにアクセスできない場合がある．また，ブックマークしたページが更新され，内容が変化している場合がある．そこで，ブックマーク時と現在のページの内容を Web アーカイブに格納された Web ページの履歴情報を用いてその変更状況を補間し，ユーザに呈示する機構を提案する．本稿では，時間によって変化するページの内容を考慮したブックマークの意味的補間方式について述べ，提案する方式に基づくプロトタイプの実装方法について検討する．

キーワード　Web アーカイブ，Web ブックマーク，情報検索

Semantic Interpolation of Bookmark using Web Archives

Tomoyo KAGE†　　Kazutoshi SUMIYA‡

† Himeji Institute of Technology, School of Humanities for Environmental Policy and Technology
1-1-12, Shinzaike Honcho Himeji-shi, Hyogo, 670-0092 Japan
‡ University of Hyogo, School of Human Science and Environment
1-1-12, Shinzaike Honcho Himeji-shi, Hyogo, 670-0092 Japan
E-mail:　†na01v072@stshse.u-hyogo.ac.jp,　‡sumiya@shse.u-hyogo.ac.jp

Abstract　Web bookmark are useful for revisiting pages which users has visited. However users can not always access the bookmarked pages because conventional bookmark consist of only the URLs. We propose a support method for users to show the additional pages, which can be understood the semantic difference between the current page and the bookmarked page by users. In this paper, we describe the semantic interpolation method based on detecting the changes of bookmarked pages. We also explain a prototype system based on our proposed method.

Keyword　Web Archive, Web Bookmark, Information retrieval

1. はじめに

多くの Web ユーザが，興味ある内容や知っておきたい情報，必要な情報などの Web ページをブックマークしている．しかし時間の経過によって様々な不都合が生じる場合がある．いつ何のためにブックマークしたかを忘れたり，ブックマークが増えると管理がしにくくなり有効に利用できなかったりする場合である．また，ブックマークが古くなるとリンクが切れアクセスできない場合や，ページの内容が変化している場合もある．我々はブックマークの不都合の一つ，Web ページの時系列変化に着眼し，Web コンテンツの変遷状況を分かりやすくするための補間方式について検討する．

例えば，旅行会社のホームページを例として挙げると，旅行会社は行く側の需要とツアー会社や空港会社などの供給側の両者から成り立ち，それらの関係によって内容が変動する場合がある．ある時点ではアメリカが人気で価格が高いプランもたくさんあり，アメリカ特集も組まれているが，次の時点ではアメリカのツアーは極端に減り，しかもかなり安値の航空券が存在する．この２つのページを見るだけでは何故このような変化が起こったかを理解することは容易ではない．何故ならその原因である当時大騒動となった事件，アメリカのテロについて旅行会社は記載していないからである．しかしその事件に旅行会社が影響を受けたことは事実で，因果関係が明確である．まだ記憶に新し

い出来事なので推論することが可能だが，これから数年，あるいはそれ以上の年月が経つと，この変化について理解することは難しい.この変化に関する情報(ここではアメリカでテロ発生）を補間情報と呼ぶ.

ブックマークしたページのリンクが切れている場合や，過去のページを見たい場合は，Web アーカイブの代表的なシステムである WayBack Machine (Internet Archive[1])を使用し過去の Web ページを取得することが可能である．WayBack Machine では URL をキーとして問い合わせると,その URL の時系列ページへのリンクが出力され，ユーザは任意の時間に存在していたページを閲覧できる．しかし変遷を知るには過去の時系列ページを見るだけでは分かりにくい場合がある．なぜなら過去のページはその当時の社会的・一般的な情報を暗黙の了解としているため，記載する必要がなく，書かれていない場合が多いのである．また，変化した要因が記載している場合でも現在の我々にとっては理解しがたい状況が充分起こりえる．そこで，ユーザが過去の Web ページを見る場合，より理解しやすくするための方法として，時系列ページの変遷（図 1）について情報を補う補間を提案する．ユーザが閲覧していない間の時系列 Web ページの履歴から内容が変化する時点を発見し，何故そのように変化したかという情報を Web アーカイブから取得して,ユーザに呈示する.

以下，2 節では関連研究について，3 節では意味的補間について述べる．4 節ではトピックキーワードについて，5 節では Web アーカイブを利用した検索について,6 節ではプロトタイプシステム,7 節ではまとめと今後の課題について述べる.

図 1　時間に伴う Web ページの変化
姫路工業大学ホームページ[16]

2. 関連研究

ブックマークに関する研究のほとんどが分類，管理の効率化を図るものである．中島らのコンテクスト依存型ブックマーク[2]ではユーザのブックマーク時の閲覧履歴をメタデータとして保持するものである．ユーザの意図が表せると共にリンク切れにも対応し，ユーザの要求に応じた代替ページの取得が可能である.

日野らの WebFarm[3]ではコンテクスト情報に加えて，ユーザのブックマークの利用頻度や支持などもメタ情報として保持し，アミューズメントの要素を取り入れた動物型の能動的なブックマークを提案している.

本研究では,ユーザの閲覧履歴を考慮せず Web アーカイブに格納されている Web ページの履歴から変遷情報を抽出し補間するものであるため，ユーザの閲覧履歴を用いるこれらの研究とは異なる.

表 1　関連研究と本研究の違い

	ユーザのブックマーク時の閲覧履歴	ユーザのブックマークの履歴	アーカイブによる Web ページの履歴
コンテクスト依存型	○	―	―
WebFarm	○	○	―
本研究	―	―	○

3. 意味的補間

3.1. 基本概念

ブックマークの意味的補間とは，ユーザがブックマークした時から現在までの時系列ページの変遷を見る時に，情報を補ってより分かりやすく呈示することである．ここでいう変遷とは，色調やフレーム，構成といったレイアウトの変化ではなく，内容の変化を意味し，内容に基づく補間という意味で“意味的補間”と呼ぶ．本研究では，内容が変化した時点を新しいトピックや新しい事象が現れた時点と考え，その変化に対する情報を補う．補う情報は，変化した時点の Web アーカイブから取得する．この情報を補間情報と呼ぶ．内容が変化した時点の過去の Web から情報を取得することにより，現在とは異なる当時の社会的・一般的情報を得ることができる．またブックマークした時系列ページを見るだけでは分からない変化を客観的に分かりやすく補間することができる.

本方式では，図 2 のようにブックマークしたページの時系列ページから内容の変化を抽出し，その時点の Web(図 2 の雲形は Web を表す)に問い合わせをして補間情報を取得する．まず，ブックマークしたページの時系列ページを取得し，その時系列ページに存在するキーワードの変遷に着目する．キーワードの変遷によってブックマークしたページの内容が変化した時点を発見する．また，Web アーカイブから補間情報を取得するために,検索式を生成する.キーワードの変遷は，出現区間と重要度によって表現し，出現区間はキーワードを含んでいる Web ページの時区間のことをいう.重要度は従来の算出法 tf-idf 法[4]を時系列 Web ページに対応させ拡張して用いる．時系列 tf-idf 法は，時系

列 Web ページにおけるキーワードの重要度を計算することが可能である．

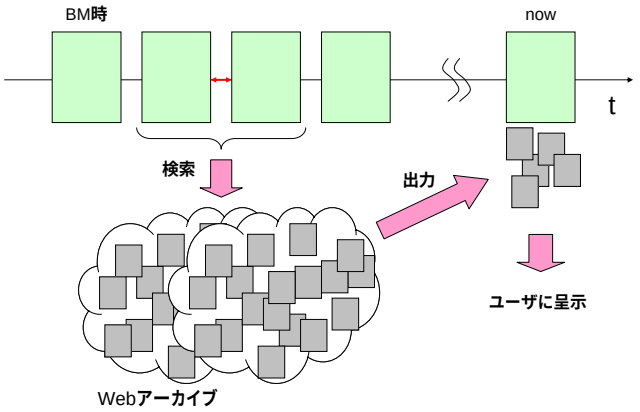

図 2　Web アーカイブを利用した意味的補間

3.2. 時系列 tf-idf 法

時系列 Web ページにおけるキーワードの重要度算出方法について説明する．一般的にキーワードの重み付けには， tf-idf 法が用いられる．しかし tf-idf 法は時間を考慮していないため，これを拡張する（図 3）．時系列ページにおけるキーワードの重要度算出は，ある時点の重要度を一時的な尺度ではなく，その前後からみた尺度を用いる．つまり時系列順序で前（過去）からみた重要度と，後（未来）からみた重要度の 2 種類の tf-idf 値を算出する．

この 2 種類の tf-idf 値の算出方法について述べる．まず，単語の頻度(tf 値)は従来の方法同様に求める．すなわちキーワードの重要度を表す 2 種類の tf-idf 値の分子は等しい．次にドキュメントの頻度(df 値)について説明する．df 値も計算の方法は従来と同様であるが，df 値の対象となるドキュメント集合が異なる．従来は tf 値を求めるページが属するドキュメント集合を用いるが，時系列 tf-idf 法では，時系列順で前のページが属する集合と後のページが属する集合を対象とする．前の集合を用いた重要度を出現度，後の集合を用いた重要度を消失度と定義する．

ブックマーク時のページをP_0，ブックマークしたページの現在のページをP_nとし，P_0からP_nまでの時系列ページを$\{P_0,P_1,\ldots P_n\}$とする．P_0からP_nの各時間に対応しているページの集合を$\{D_0,D_1,\ldots,D_n\}$とし，任意のページ$P_i(i \geqq 1)$におけるキーワードkjの出現度をap_{kij}，消失度を$disap_{kij}$とすると，2 つの重要度は以下の通りである．

$$ap_{kij} = tf_{ij} \cdot idf_{(i-1)j} = tf_{ij} \cdot \log \frac{N_{i-1}}{df_{(i-1)j} + 0.1} \quad (1)$$

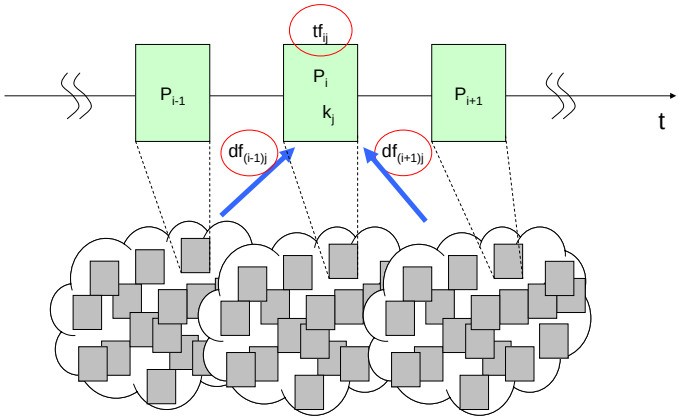

図 3　時系列 tf-idf 法

$$disap_{kij} = tf_{ij} \cdot idf_{(i+1)j} = tf_{ij} \cdot \log \frac{N_{i+1}}{df_{(i+1)j} + 0.1} \quad (2)$$

分母をdf+0.1 としているのは，値が 0 になることを避けるためである．従来法ではP_iが含まれているドキュメント集合を用いて算出するため，df値が 0 になることはない．しかし，時系列tf-idf法では，Piが含まれていないドキュメント集合を用いるため，分母が 0 になる場合がある．

この 2 種類の出現度と消失度の差異を重要度の変化量とする． WebページP_iにおけるキーワードk_jの重要度の変化量は以下の通りである．

$$w_{ij} = ap_{kij} - disap_{kij} \quad (3)$$

出現度が高いキーワード，すなわち新しく出現し，増加傾向にあるキーワードは正の値となり，消失度が高い値のキーワード，すなわち重要度が下がり減少傾向にあるキーワードは負の値となる．また，重要度が変化しない場合は 0 となる．

4. トピックキーワード

Web ページは複数のトピックを扱っているため，どのトピックが出現，消失，変化したかといった変遷を明確にすることは容易でない．そこで本研究では新しいトピックやトピックの推移を抽出するために，時系列ページ中の複数のキーワードをトピック毎に分類する．まず，4.1 節では変遷に影響を及ぼさないキーワードの除去を行い，4.2 節でクラスタリングする．

4.1. フィルタリング

Web ページの変化に影響しないキーワードをフィルタリングする．つまり Web ページの変遷に関わりがないキーワードを除去する．フィルタリングの方法は以下の 3 つである．図 4 と表 2 の例を用いて説明する．

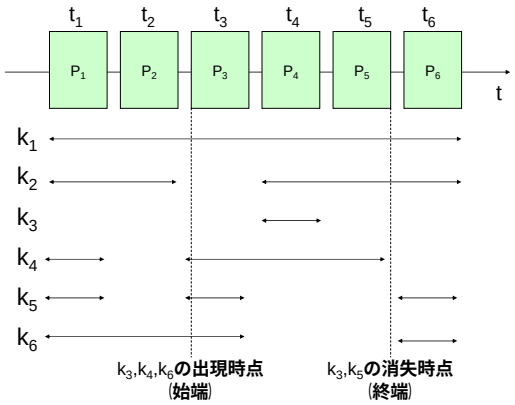

図 4　キーワードの出現区間

- 常時出現しているキーワード（図 4 のk_1）は，Web ページを構成している基本的なキーワード，または一般的によく用いられるキーワードであるため，Webページを変化させる要因にはならないと見なし，除去する． 時系列Webページの最初と最後のページに出現し，ほとんどのページに出現しているキーワード（図 4 のk_2）もWebページを構成している基本的なキーワードとしてフィルタリングする．ここで閾値は時系列Webページの全時区間の 0.8 とする．
- 一時的に出現しているキーワード（図 4 のk_3）は，Webページの内容を変化させるには時間が短く不十分で，Webページに影響を与えないとしてフィルタリングする．一時的な出現区間が複数ある場合（図 4 のk_5）やある程度連続して出現している区間が別に存在する場合（図 4 のk_4, k_6）は，一時的な出現とはみなさない．一時的とは時系列Webページ全区間において，0.05 以下の時区間をいう．
- 新たに出現しているにもかかわらず，出現時点の重要度の変化量が負である場合，そのキーワードは特徴的なトピックにはならないと見なしフィルタリングする（表 2 のk_2）．また，出現区間が途切れている場合は，連続した出現区間のその始端を出現時点と考える（図 4 のk_4のt_3，k_6のt_1の特徴量）．なお，一時的な出現は考慮しないので，連続した出現区間の始端の値が全て負である場合，一時的な出現区間の値が正でもフィルタリングする．例えば，表 2 のk_6はフィルタリングの対象となる．図 4 のk_5は一時的な出現区間のみで構成されるため，例外と考えフィルタリングしない．

このように 3 つの観点によるフィルタリングを行うことによって，Web ページの変遷に関係するトピックキーワードを抽出する．

表 2　キーワードの特徴量

	t_1	t_2	t_3	t_4	t_5	t_6
k_1	+0.1	+0.0	-0.2	-0.3	+0.2	+1.2
k_2	-0.2	+0.4		+0.1	-0.8	-0.0
k_3				+2.5		
k_4	+2.0		+0.1	-0.3	+0.2	
k_5	+1.0		+1.3			-0.2
k_6	-1.4	-1.5	-0.1			+1.0

4.2. クラスタリング

フィルタリングしたキーワードをトピック毎にクラスタリングする．クラスタリングは,「同一のトピックであるキーワードは類似した変遷である」と考え，出現区間及び重要度の変化量に基づいて行う．4.2.1 節では出現区間について, 4.2.2 節では重要度の変化量についてのクラスタリング手法を説明する．

4.2.1. 出現区間

最初に，出現区間の情報のみでクラスタリングを行う．出現区間が 1 つしかない場合と出現区間が途切れて複数ある場合が考えられるが，まずは前者について述べる．

出現区間が 1 つの場合，基本的には同一の区間でクラスタリングする．ただし，クラスタを構成するキーワード数が少ない，または同一の出現区間であるキーワードが存在しない場合は，誤差率 20%を許してクラスタリングする．出現区間が同一でないキーワードのクラスタリングは，クラスタできないキーワードが優先され，次にクラスタ内のキーワードが少ない順に優先される．どのクラスタにも該当せず，誤差率の大きい場合，クラスタリングは行わない．

次に，出現区間が複数ある場合のクラスタリングについて説明する．

(1) 連続した出現区間が同一のキーワードをクラスタリングする．全時区間ではなく，出現区間毎にクラスタリングを行う（図 5）．この時，異なるクラスタに同一のキーワードが含まれる形を許す．

(2) (1)の処理後，出現区間の時区間の幅を広げ，同様のクラスタに属するキーワードを更にクラスタリングしていく（図 6）．方法は始端と終端が共通している区間を拡張していき，出現区間を全て含む区間まで行う．

(3) 同様の出現区間であるクラスタ（誤差率 0%）を構成しているキーワード数が少ない場合は，条件を緩和し，異なる出現区間を含んでいても構わないとしてクラスタ及びキーワードを結

合する．ただし，出現区間が 1 つの場合同様，誤差率を全時区間の 20%とし，キーワードが少ないクラスタが優先される．

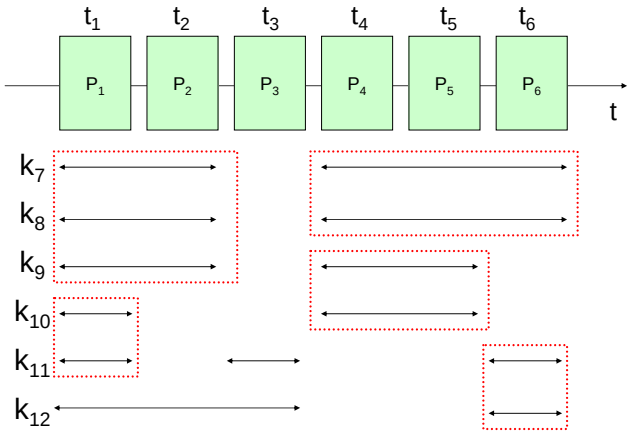

図 5　出現区間のクラスタリング

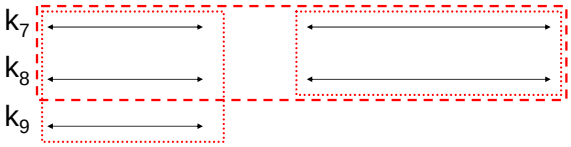

図 6　出現区間を拡張したクラスタリング

4.2.2. 重要度の変化量

出現・非出現の関係のみではトピックが定まらないので，更に重要度の変化量を考慮し，同一の内容に関するキーワードを分類する．クラスタリング方法は，重要度の変化量が正か負の値に分類し，同じパターンのキーワードを同じクラスタとする．前節のクラスタリング時に少ないキーワードで構成されたキーワードは，ほぼ同様の変遷であると考えられるためクラスタリングしない．また，前節と同様にクラスタを構成しているキーワード数が少ない場合のみ誤差を許す(85%).生成したクラスタ内のキーワードをトピックキーワードとする．

5. Web アーカイブを利用した検索

4 節で述べた方法で，時系列 Web ページが変化した時点を発見し，変化した時点の Web に問い合わせを行う．Web アーカイブに時間指定し，過去のある時点の Web から補間となる情報を取得する．時間を特定することによって，その時間特有の，タイムリーで有用な情報が抽出できる．また，現在ではあまり特徴的でないキーワードでも当時話題となっている事柄や情報に関連していれば，そのキーワードによって社会的な情報の取得が可能になる．時間を指定するために，まず Web ページの内容が変化した時点を発見する．本節では，「新キーワードの出現による Web ページの変化」と「トピックの推移による変化」の 2 種類の内容の変化を検討する．2 つのパターンについてそれぞれ変化した時点，補間情報の抽出方法を説明する．

5.1. 時間の特定

5.1.1. 新キーワードの出現

時系列Webページに新しいトピックが出現した時点を特徴的な変化と見なす．新しいトピックの出現を，重要度の高いキーワードの出現と考え，キーワードk_jの出現時に，重要度の変化量w_{ij}が他と比べて極めて大きい場合，その時点を変化した時点とする．

5.1.2. トピックの推移

トピックの推移とは，時系列ページ中に異なる時間に出現している関連したトピックのことをいう．本稿では時系列で前後したトピック 2 つを扱う．なお，前者と後者は互いに影響を与えているものとする．

あるトピックの出現後に異なるトピックが出現すれば，影響を及ぼしあっていると判断できるため，トピックキーワードの出現区間が接している場合（図 7 C_1とC_2），両者は推移していると考え，その接している時点を変化した時点とする．ここで重複する場合(図 7　C_1とC_3，C_1とC_4など）も同様に考える．ただし，図 7 のC_1とC_5のように完全な包含関係の場合，前後関係が判断できないので除く．

旅行会社のホームページの例では，ある過去の時点に『海外旅行人気』,『お正月は海外で！』といった海外ブームの内容が記載されている．また,『アジア』の推薦もしている．しかし次のページではアジアの国名は一気に減少し,『国内旅行』や『温泉』の内容が目立っている．ここで挙げた例では，ある時点までアジアに関するトピックが出現した後，国内に関するトピックが出現している．この時点がトピックの推移した点である．

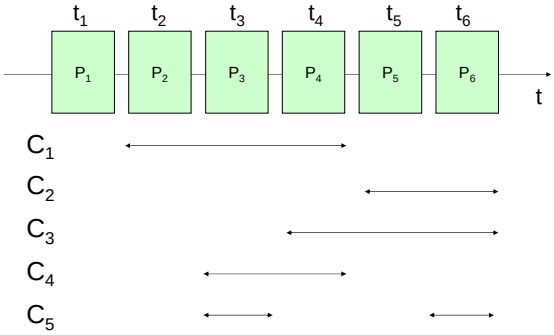

図 7　クラスタの時間関係図

5.1.3. 周期的変化

変化がシーズン毎や，月毎に現れることがある．たとえば，旅行会社のホームページでは寒い時期になるとスキーやスノーボードなどのゲレンデ情報が一気に

増える．このような変化は，社会的な変動とは異なる周期的な変化である．ただし，本稿ではこの場合については紙面の都合上考慮していない．

5.2. Web アーカイブの検索

5.2.1. 新キーワードの出現

新キーワードの出現による変化の場合，検索対象となる Web は重要度の変化量が高いキーワードが出現した時点である．キーワード質問は，その変化量が大きなキーワードを含めたトピックキーワードによって生成する．類似した新キーワードが複数出現する場合は，新キーワードのみを用いてキーワード質問を生成する．新キーワードが 1 つ，またはあまり類似していない新キーワードの出現であるなら，トピックキーワードを補って質問生成する．補うキーワードは，新キーワードに類似しているとし，情報を特定するため AND 条件とする．しかし全て AND 条件で問い合わせてしまうと，トピックキーワードが異なるトピックであった場合，補間ページが取得しにくい．そこで，新キーワードの変遷と類似しているキーワード上位 2 つを抽出して，新キーワードと類似キーワードを AND 条件で括り，2 つのキーワード質問とする．検索式のうち多くヒットしたページが補間ページの候補となり，そのうち質問となっているキーワードが特徴的なページを補間ページとする．

5.2.2. トピック推移

トピックの推移による変化の場合，両者の関係を分かりやすくするため，両者のキーワードが記載しているページすなわち両者に関する情報が記載されているページを補間ページとして取得する．そこでキーワードが共に出現しているページを取得する．通常，キーワードが共起しているページを取得する場合はキーワードとキーワードを AND 条件で括って検索式を作成する． しかし AND 条件ではキーワードが共起している“ある時点”のページしか取得することができない．Web アーカイブは同一の時系列ページを複数格納してあるので「同一のページ」＝「同一の時系列ページ」となり，時間を越えてキーワードが共起しているページを取得することができる．そこで通常の AND 検索では時系列で共起したページが取得できないので，検索式を緩和して問合せを行う．前に出現したトピックの終端である時点と，後に出現したトピックの始端である時点の Web に，キーワードを AND 条件ではなく OR 条件で問い合わせる．前のトピックキーワードと，後のトピックキーワードのそれぞれ重要度が高い上位 m 件（ここでは 2 件）により生成する．

旅行会社のホームページの例では，前に出現したトピックはアジア，後に出現したトピックは国内で，トピックキーワードは，前者が『中国』や『香港』，後者が『国内』，『温泉』とする．現時点では実際の Web アーカイブを用いた検索ができないが，SARS の影響で海外旅行会社や航空会社の市場が赤字であるといった内容のニュース記事や市場に関するページが補間ページであると想定できる．この補間ページによって何故旅行のホームページからアジアの国々が急激になくなったのかを理解することができる．実はこの例では『中国』が検索式のキーとなり，この時点の Web アーカイブに『中国』で検索すると『SARS』のページをたくさん取得できると考えられる．しかし，変化が起こった原因と関わりが深いキーワードが何であるのかは分からないので，先ほど述べたように網羅的に様々なキーワードの組み合わせを考慮して問い合わせる必要がある．

5.3. 補間ページの呈示

補間ページは何故ブックマークしたページが変化したかについて情報が記載されているページであると考えられる．因果関係のあるページの変化は，時系列に閲覧した方が理解し易い．本研究では，ブックマークしたページの変遷に加えて，変化時点に補間ページを挿入して呈示する(図 9)．呈示はスライドショーのような方法，あるいは補間ページ中の画像を利用した時系列ページのモーフィング提示など視覚的な呈示方法が考えられる．また，テキスト情報として新キーワードの強調も考えられる．

現在，レイアウトの変更によるページの変化も一緒に見られるユーザインタフェースを検討中である．

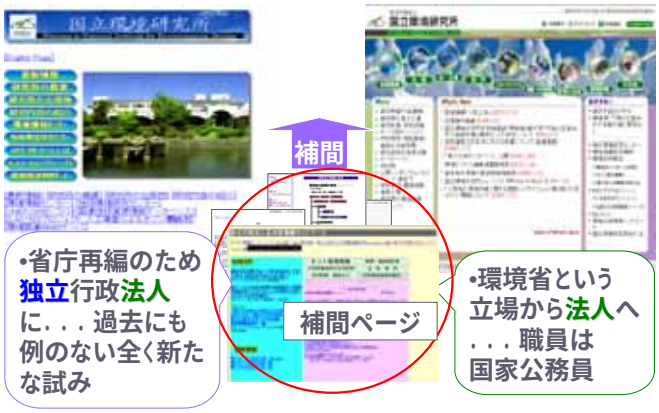

図 9　　補間の呈示例
独立行政法人国立環境研究所のホームページ[14]
左図 1997 年　右図 2005 年

6. プロトタイプシステム

6.1. 実装

プロトタイプは C#を用いて構築した（フィルタリング，クラスタリング，クエリー生成などは一部手作業を含む）．以下手順について説明する（図 10）．

(1) ブックマークの URL をキーとして Internet Archive から時系列 Web ページを取得する．取得方法は，(http://web.archive.org/web/*/ 任意の URL)にアクセスし，ページのリンク情報を抽出する．

(2) 時系列 Web ページと同時にそのサイト内の Web ページを取得する．HTML ドキュメントのみを対象とし，他のドキュメントは取得しない．なお，3 階層のサイト内のページ数は数十から数百ページである．

(3) tf-idf 値の算出方法については，形態素解析ソフト「茶筅」[15]を用いて単語（名詞）抽出し，名詞のみに対して行う．ブックマークした時系列 Web ページに対して tf，サイト内のページに対して df を求めて計算する．Web ページにおける計算時間は数十秒程である．

(4) 抽出したキーワードに対して，時系列 Web ページ毎の特徴量を計算する．

(5) キーワードの出現・非出現区間，特徴量値の正負により，キーワードをクラスタリングする．

(6) クラスタの出現区間・特徴量により，変化を発見する．

(7) 変化に関するクラスタから特徴量の大きいキーワードを用いてキーワード質問を生成する．

(8) 検索は，Web ページをテキストファイルの形式で読み込み，該当するキーワードを含むページを取得する．

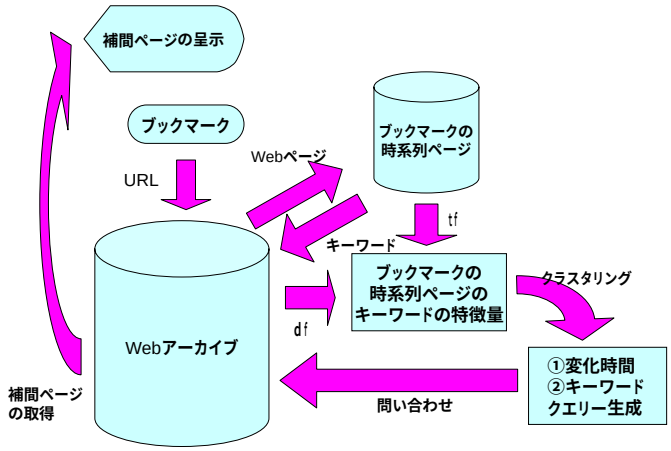

図 10　システム構成図

6.2. 実験と考察

文部科学省[12]，東京大学[13]，独立行政法人国立環境研究所[14]の 3 件の Web ページをブックマークページとして実験を行った．3 件は，時系列 Web ページ全体の時間幅，ページ同士の時間幅を違う形式で取得した．文部科学省では 2001 年の 2 月から 5 月の連続した 7 ページ，東京大学では 2000 年から 2003 年の 4 ヶ月毎の 11 ページ，独立行政法人国立環境研究所では 1998 年から 2004 年の 1 ヵ月毎の 26 ページの時系列 Web ページを取得して実験を行った．

なお，時系列ページの時間幅はインターネットアーカイブの収集状況に左右され若干揺れがあるが，ここでは考慮しないものとする．

6.2.1. フィルタリング

時間の変化に影響されない，また影響しない単語の除去のためのフィルタリングの有効性について実験を行った．その結果，名詞の数が，文部科学省のページでは 75 から 37 に，東京大学のページでは 193 から 80 に，環境研究所のページでは 573 から 249 に削減した．

- 全区間に出現していたため削除されたキーワードは，文部科学省では『教育』,『科学』,『文部』など，東京大学では『事務』,『問い合わせ』,『お知らせ』など，環境研究所では『国立』,『環境』,『研究所』などである．
- 一時的に出現していたため削除されたキーワードは，文部科学省では『宇宙』,『ロシア』など，東京大学では『キャンプ』,『都市』など，環境研究所では『来訪』,『衆議院』などである．
- 特徴的でないため削除されたキーワードは，文部科学省では『年』,『週間』など，東京大学では『地区』,『学生』,『本』など，環境研究所では『ガイド』,『比較』などである．

表 3　フィルタリングの例

	文部科学省	東京大学	環境研究所
全区間	教育 科学 文部	事務 問い合わせ お知らせ	国立 環境 研究所
一時的	宇宙 ロシア	キャンプ 都市	来訪 衆議院
低特徴	年 週間	地区 学生 本	ガイド 比較

名詞除去を行うフィルタリングでは，いずれも 50%の単語の削減を実現した．削除したキーワードもほぼ妥当な結果となり，フィルタリングが有効であることが分かった．

6.2.2. 新キーワードの出現

新キーワードの出現に基づく補間方式の有効性について東京大学のデータを用いて実験を行った．

新トピックが 5 件見られた．うち解となったのは 4 件で，『パブリック』,『コメント』や『公開』といったキーワードであった．『パブリックコメント』は，2001

年 8 月に特徴的なキーワードで，この時，東京大学が大学憲章を定めることになり，そのパブリックコメントが出現したと思われる．また，『公開』が出現したのは 2001 年 4 月で，その年の 10 月に施行される「独立行政法人等の保有する情報の公開に関する法律（誰でも東京大学の保有する法人文書の開示を請求することができるという法律)」に先立って，行政文書の情報公開に関するトピックが出現したと思われる．

このように，特徴的なキーワードの出現における新トピックが抽出された．

6.2.3. トピックの推移

トピックの推移に基づく補間方式の有効性について東京大学のデータを用いて実験を行った．

トピックが重複するパターンが 6 件見つかり，うちトピックの推移であったのは 2 件であった．解の 1 つの前方のトピックキーワードは『パブリック』,『コメント』で，後方のトピックキーワードは『UT』,『会議』であった．これは，大学憲章に関するパブリックコメントを UT21 会議がまとめ，論点整理案を公表したという関係があることが分かった．

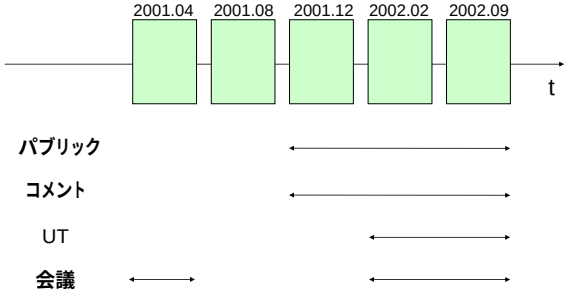

図 11　キーワードの出現区間（東京大学の例）

表 4　キーワードの重要度変化（東京大学の例）

年月 / キーワード	01.04	01.08	01.12	02.02	02.09
パブリック			+2.69	+0.30	+0.37
コメント			+2.69	+0.30	+0.37
UT				1.74	-0.24
会議	-0.66			+0.59	-0.15

7. まとめと今後の課題

本論文では，Web アーカイブを利用して，ブックマークした Web ページの変遷について補間する方式を提案した．現在から過去のページを見る場合に，当時の情報が必要不可欠である，という視点から，Web アーカイブを利用して補間情報を取得する方法について検討した．ブックマークの新しい機能として有効であるだけでなく，Web アーカイブの新しい活用方法であると考えられる．

今後の課題として実験に用いるデータ数を増やすこと，また，ユーザに対して時系列 Web ページの変化を分かりやすく呈示するユーザインタフェースの開発が挙げられる．

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] Internet Archive: Way Back Machine, http://www.archive.org/

[2] 中島伸介，黒田慎介，田中克己：閲覧履歴を反映したコンテクスト依存型 Web ブックマーク．情報処理学会論文誌：データベース，Vol.43, No.SIG5 TOD14, pp.23-36(2002)

[3] 日野洋一郎，中島伸介，小山聡，田中克己：WebFarm:動物メタファを用いた Web ブックマーク再利用機構，第１４回データ工学ワークショップ DEWS’03, (2003.3.3-5)

[4] 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)

[5] 角谷和俊，田中克己：Web アーカイブのための時間情報管理とその応用，情報処理学会研究報告 DBS-131, pp.109-116 (2003)

[6] 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, (2003.3.3-5)

[7] WARP:http://warp.ndl.go.jp/

[8] VisualMarks1.2:http://www.6bytes.com/visualmarks.html

[9] NetXtract:http://www.netxtract.com/

[10] http://www.informatik.uni-trier.de/~ley/db/conf/sigmod/LiVCAHMWBCHIKSS99.html

[11] Cyclone:http://cyclone.slis.tsukuba.ac.jp/

[12] 文部科学省: http://www.mext.go.jp/

[13] 東京大学: http://www.u-tokyo.ac.jp/index-j.html

[14] 独立行政法人国立環境研究所:http://www.nies.go.j/index-j.html

[15] 茶筅: http://chasen.aist-nara.ac.jp/hiki/ChaSen/

[16] 姫路工業大学: http://www.himeji-tech.ac.jp/

[17] 豊田正史，喜連川優：大規模 Web アーカイブからのデータマイニング，情報処理，vol46,No.1,pp.46-51 (2005)

[18] 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，人工知能学会研究会資料 SIG-SW&ONT-A401-01