Webからの時制クラスタの解釈

Labeling Temporal Cluster of Web Pages

森 正輝♥　三浦 孝夫♦
塩谷 勇♠

Masaki MORI　Takao MIURA
Isamu SHIOYA

本稿では，Webページ集合からの事象抽出及び自動解釈を行うための新しいWebマイニングの手法を提案する．まずWebページから有効時間を推定し，K-meansクラスタリングにより事象の抽出を行う．更に，各クラスタを解釈するため，KeyGraphを用いてラベル付けを行う．提案する手法が時制Webページに対して有効であることを実験により示す．

In this investigation, we propose a new mechanism of Web Mining to extract events and generate labels from a collection of Web Pages. Here we examine Web pages and obtain valid timestamp, and detect events by means of clustering. We generate labels based on KeyGraph technique and show how well our approach works to temporal Web pages by some experiments.

1. 動機と背景

近年のWebページの総量は莫大なものであり，驚異的なスピードで増え続けている．この情報洪水の状況で，利用者はWebページ集合が何を表しているか理解することが難しくなる一方であり，Webページ集合の内容を素早く容易に把握するための研究が近年注目を浴びている[2][9][10].

現在，Google，Yahoo!等の検索エンジンを使えば，利用者は適切な検索語を与えることで，いくつかのトピックを得ることができる．利用者にとって望ましい情報を見つけるため，多くの検索エンジンは 3億から30億と言われる巨大なURLデータベースを構築している．これにより探索の手間を軽減させることができるが，検索結果のリストが長くなるという新たな問題が発生する．利用者は検索結果をブラウズし有益なWebページを探すが，多くの場合，途中で断念してしまう．

現在，参照の数，ハブとリンクなどのオーソリティ値，個人の好みなどの統計的な値を用いるなどの手法が提案されている[5]．しかし，リストが示す内容を一見しただけで理解するのは困難であり，どれほどうまく並べられても，どのような事象が起きているかを理解することは難しい．解決法の１つとしては，ページを意味的にグループ化することが考えられる[4]．検索したページをクラスタに分類し情報を要約できたならば，検索結果をより効果的に容易に吟味することができ，利用者の負担も軽減されると考えられる．

ページの有効時間を推定することができれば，これを手がかりとして内容を事象ごとに理解することができ，Webページから時間軸上で自動的に事象を抽出することも可能になる．この一連のアプローチをTopic Detection and Tracking (TDT)と呼ぶ[2][8].

本稿では，はじめに各Webページのタイムスタンプの推定を行う．通常，文書はその内容に関する時間，すなわち有効時間(valid time, VT)に従って理解されるが，必ずしも文書の内容時間(content time, CT)が文書の作成・修正された時間，すなわち作成時間(creation time, UT)やトランザクション時間(transaction time, TT)と一致するものではない．有効時間を推定しWebページを時間軸に従ってクラスタ化すれば，個々のクラスタは意味のある事象に対応すると考えられる．更に，個々のクラスタの意味解釈を自動的に与えるため，KeyGraphに基づく手法を用い，クラスタから重要語を抽出し，これをクラスタラベルとして付与する手法を提案する．

本稿では，２章でWebページから有効時間の推定，３章で事象の抽出方法とその有効性を論じ，Googleを使った実験的な結果を論じる．４章では，事象を解釈するラベルの決定方法を論じ，５章で実験結果の考察を行い，６章で結論とする．

2. 有効時間の推定

Webページから有効時間を推定するために，本稿では３種類の時間を考慮する．即ち，コンテンツに明示的に現れている**内容時間**，**作成時間**，**更新時間**である．作成時間はURLの一部に含まれおり，更新時間は受信ヘッダに明記してあることが多い．有効時間 VTとは，Webページが示そうとしている時間を意味する．内容時間 CTとは，Webページ文書に明示的に出現しているタイムスタンプである．これは各文書の最初の文に現れ，「Jan 04, 2004」又は「January 3, 2004」のような一定のパターンを想定する．複数の内容時間が抽出できる場合はすべてを扱う．例えば「松井稼頭央が2003/12/09 にNew York Mets に入団」という文では，「2003/12/09」がこれに該当する．作成時間 UT とは，Webページが生成された時間を言う．次の例が示すように，経験的に作成時間はURLの一部として現れることが多い．

http://dsc.discovery.com/news/afp/20040105/marspix.html
http://www.cbsnews.com/stories/2004/01/04/tech/main591195.shtm

それぞれのURLは2003/01/05，2004/01/04 の作成時間を含んでいる．このとき，Webページの著者は原稿にしたがってページを生成するので内容時間と作成時間は必ずしも等しくならない．更新時間 TT とは，Webページが格納された時間または最後に更新された時間を言う．更新時間については，受信ヘッダファイルに"Last-Modified:Tue,19 Aug 2003 06:10:54 GMT"のようなLast-Modifiedのヘッダが含まれる場合，そのWebページが「2003/08/19/06:10:54」で格納されたか，あるいは最後に更新されたことを意味する．

各Webページを解析し，内容時間，作成時間，更新時間を抽出し，どれが有効時間に近いかを推定する．無論，すべてのWebページに CT，UT，TTを必ず含むわけではない．以下では，抽出したCT，UT，TTのどれがVTに近いかを調べるため，テキストコレクションを取得し，手作業によりそれぞれのページの有効時間を調べる．このとき，CTがなくとも UTやTTを持つ場合がある．時間が入手できないことを"null ”を用いて表すとし，テストページの集合Tを取得す

♥ 学生会員　法政大学大学院工学研究科電気工学専攻
i04r3246@k.hosei.ac.jp
♦ 正会員　法政大学工学研究科電気工学専攻
miurat@k.hosei.ac.jp
♠ 正会員　産能大学経営情報学部情報学科
shioya@mi.sanno.ac.jp

る：

$T = \{< p_i, VT(p_i), ET(p_i), CT(p_i), UT(p_i), TT(p_i) > | i = 1, 2, ...\}$

ここで $T_P = \{p | < p, .. >\in T\}$ と定義する．$p \in Tp$ が与えられたとき，その推定時間ET(p) を得るため，V, P_C, P_U, P_T, P_{CU}, P_{UC}, P_{CT}, P_{TC}, P_{UT}, P_{TU}, P_{CTU}, P_{CUT}, P_{TCU}, P_{TUC}, P_{UCT}, P_{UTC} を次のように定義する：

$V = \{< p, VT(p) > | p \in TP, VT(p) \neq null\}$

$P_C = \{< p, CT(p) > | p \in T_P, CT(p) \neq null\}$

...

$P_{CT} = P_C \cup \{< p, TT(p) > | p \in T_P, CT(p) = null, TT(p) \neq null\}$

...

$P_{CTU} = P_{CT} \cup \{< p, UT(p) > | p \in T_P, CT(p) = null, TT(p) = null, UT(p)\ 6= null\}$

...

集合Vは全ての可能な推定時間を意味する．P_C, P_U 等の定義はどのように推定時間(ET)を得るかを示す．例えば，P_{CU} はCTがnullでない限り有効時間として，CT(p)がnullだがUT(p)がnullでないときはUT(p)を有効時間として推定する．この意味で，P_{CU}は有効時間の推定方法を示しており，以下ではこれをCUと示す．Ans(答)，Rec(再現率)，Pre(適合率)，F値を次に定義する：

$Ans(P) = \{< p, t >\in P \mid t = VT(p), t\ 6= null\}$

$Rec = |Ans(P)|/|V|, Pre = |Ans(P)|/|P|$

$F = 2 \times Rec \times Pre/(Rec + Pre)$

すべての組み合わせでF値を算出し，実験的に最大のF値のもの選択する．これを決定すれば，Webページから推定時間を得る方法を求めたことになる．

表1　有効時間の推定

Table 1 Estimating Valid Time

Scheme	ExpTime	Ans	Pre	Rec	F
C	164	127	77.4	60.2	67.7
U	52	42	80.8	19.9	31.9
T	68	2	2.9	0.9	1.4
CU	177	133	75.1	63.0	68.6
CUT	213	133	62.4	63.0	62.7
CT	202	127	62.0	63.0	62.5
CTU	213	132	62.0	63.0	62.5
UC	169	130	76.9	61.6	68.4
UCT	205	130	63.4	61.6	62.5
UT	113	43	38.1	20.4	26.6
UTC	184	105	57.1	49.8	53.2
TC	192	102	53.1	48.3	50.6
TCU	203	105	51.7	49.3	50.5
TU	113	38	33.6	18.0	23.5
TUC	184	93	50.5	44.1	47.1

本実験では，Googleに検索語"Kazuo Matsui"を与えて検索しTop300ページを得る．検索結果からリンク切れ，Weblog以外の235のURLを対象として手動でタイムスタンプを推定し，211ページのタイムスタンプを得た．同時にこれらから内容時間CT，作成時間UT，更新時間TTの抽出を行う．上記表で Schemeは有効時間の推定方法，ExpTimeはnullではない時間を持つページの数を，Ansはスキーマごとの答の数を示す．結果よりF値が最大となるスキーマはCUである．

3. 事象の抽出

TDTの分野において，時間軸におけるクラスタ化が効果的であるとはよく知られている[2]．すなわち，各トピックに対して事象(Event)は時制クラスタに対応することが多い．

ここでは時間軸でクラスタ化することの正当性を，Kazuo Matsuiのページを用いて示す．235のWebページから最もF値の高かったCUスキーマにより取得した177のページに対し，K-meansアルゴリズムを利用してクラスタ化を行う．この結果を図1に示す．ここでPageはクラスタ内のページの数を示し，CTは内容時間のページの数，UTは，作成時間のページの数を示す．結果として，8つのクラスタを得た．このうち半数は要素数5以下であり無視する．残る4つのクラスタ

Group	Time Interval	Pages	CT	UT
Group0	1975/10/23 - 1975/10/23	5	5	0
Group1	1995/06/20 - 1997/11/18	4	4	0
Group2	2000/06/27 - 2001/11/04	4	2	2
Group3	2002/03/16 - 2003/01/01	5	5	0
Group4	2003/06/29 - 2003/12/20	93	87	6
Group5	2003/12/27 - 2004/01/26	24	22	2
Group6	2004/01/31 - 2004/02/17	17	15	2
Group7	2004/02/19 - 2004/03/06	25	24	1
(total)		177	164	13

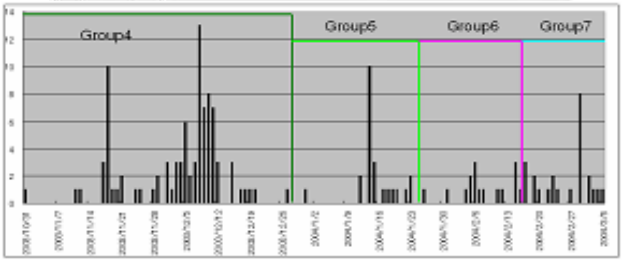

図1 Kazuo Matsui のクラスタリング結果

Fig.1 Clustering Kazuo Matsui pages

を解釈するために，Kazuo Matsuiを含む文書を選択して，頻繁な語，特徴的な語を手作業で取り出した．93ページからなるGroup4のうちいくつかの文章を例として示す．

http://www.bayarea.com/mld/cctimes/sports/7289763.htm

Seibu Lions shortstop Kazuo Matsui wants to play in the majorleagues, the seven-time Japanese League All-Star said Monday.

http://www.boston.com/sports/baseball/articles/2003/12/09/kaz_matsui_signs_on_with_the_mets/

Kaz Matsui signs on with the Mets

これらの文章より，このクラスタを"Mets welcome sign major league"と解釈した．同様にすべてのクラスタ解釈を下に示す．

(Group4: 2003/6/29 - 2003/12/20)
Mets welcome sign major league

(Group5: 2003/12/27 - 2004/1/26)
ready challenge

(Group6: 2004/1/31 - 2004/2/1 spring
opening exhibition game

(Group7: 2004/2/19 - 2004/3/6)
injure finger champ

実際に生じた事件と照合すると，すべてのクラスタの解釈はKazuo Matsuiに関して適当なクラスタであり，この事象設定は妥当であると言える．すなわち，Webページから事象を抽出するのに提案手法は有用であり，時制的な側面を持つWebページにもTDTと同じ傾向を持つことを示している．

4. 事象の解釈

本稿では，各クラスタから重要語を抽出し，これをラベルとして付与する手法を提案する．ここではKeyGraph[7]の考え方を用いる．KeyGraphとは，文書中に出現する単語の出

現頻度と共起関係から文書の主張点を把握し，重要語を抽出する手法である．KeyGraphでは，文書には必ず主張すべきポイントがあり，これらは文中に頻繁に出現する基本的な概念を用いて構築される，という仮定を設ける．基本概念とは頻出する語句であり，共起する場合にはこれらをまとめてクラスタ化する．文書中に出現する語句で，できるだけ多くの基本概念に共起するものを重要語と呼ぶ．重要語こそが筆者の主張であり，この文書の目的でもある．多くの実験例あるいは実例によって，抽出された重要語句が筆者の主張を的確に表現していることが知られている．

KeyGraphの生成はいくつかのステップからなる[7]．文書Dから不要語処理・ステミング処理を行って得た語集合Wから，上位定数個の頻出語 $w_1,..,w_N$ を抽出してその共起度を計算する[3]．すなわち，文sごとに語w_iとw_jの出現回数$|w_i|$を求め，語の共起度$co(w_i,w_j)$をのこれらの積和と定義する．

$$co(w_i, w_j) = \Sigma_{s\in D} |w_i|_s \times |w_j|_s$$

頻出語をノード，一定値以上の共起度(経験的に30)を持つノード間に辺をもつグラフGをつくり，Gの極大連結成分を土台(foundation)と定義する．土台とは頻出語を共起度でクラスタ化した語集合であり，公知概念の集合体（基礎概念）に対応するとみなすことができる．つぎにWの語wに対して，その重要度key(w) を，全ての土台概念と共起するほど1.0に近づく値として導入する．このため$|w|_s$を文sでのwの出現頻度, 土台g に対して$|g|_s$をsとgの双方に生じる語の数とする．さらに$|g-w|_s$をw∈gならば$|g|_s-|w|_s$さもなければ$|g|_s$と定義する．ふたつの関数*based(w, g)*,*neighbor(g)*を次で与える：

$$based(w, g) = \Sigma_{s\in D} |w|_s \times |g - w|_s$$
$$neighbors(g) = \Sigma_{s\in D, w\in s} |w|_s \times |g - w|_s$$

関数*based(w, g)*はgの語が生じる文でwが共起する数を，*neighbor(g)*はgの語が生じる文に含まれる語の数をあらわす．このとき*key*(w)を，全ての土台を用いるときにwを利用する条件確率とする．すなわち，

$$key(w) = probability(w | \cap_{g\subset G} g) \text{ つまり，}$$
$$key(w) = 1 - \Pi_{g\subset G}(1 - based(w, g)/neighbor(g))$$

ここで*based(w,g)/neighbor(g)*は土台gを用いるときに語wを同時に用いる割合を示している．これは土台となる語との共起度を示し，高い値を持つものを重要語とする．

本稿では，各Webページを文とみなし，KeyGraphにより時制クラスタから重要語を抽出しクラスタの解釈に用いる．実験に用いるWebページは同一トピックを論じたものであるため，得られたクラスタは相互に類似性が高く，重要語には極端な差異は生じない．一方，時間軸に沿って変化しているときには，長期的な概念も短期的な概念も含まれる．このため，「時制クラスタの解釈」を「短期的概念の変化状況の記述」と考え，直前の時制クラスタにおける重要語集合の差分と解釈する．本稿では，重要語として得られた全ての語の，上位7パーセントを差分対象に用いる．

5. 実験

5.1 実験の概要

本実験では，検索語“Hussein”という条件の下に Google より 1000 個の URL リストを取得し，リンク切れ，Weblog，時間情報のないページを取り除いた結果から事象の抽出と自動解釈を試みる．有効時間を推定しクラスタリングすることによって事象がうまく構築できることをみるため，各クラスタのタイトルおよび文で検索語を含むものを取り出し，クラスタの時間幅内に生じた事実とつき合わせて対応関係を評価する．また，これらから特徴的なラベルを人手により付与し，自動解釈により得たラベルと矛盾がないことを確認する方法で評価する．

5.2 事象の抽出

検索語“Hussein”に対してGoogleより得た1000個のURLリストから最終的に669ページを得る．これらを時間軸によりクラスタ化すると，図２のように６つのクラスタを得る．

GroupID	Pages	ContentTime	URL Time
Group0	82	75	7
Group1	101	79	22
Group2	162	129	33
Group3	57	51	6
Group4	182	156	26
Group5	85	80	5
Total	669	570	99

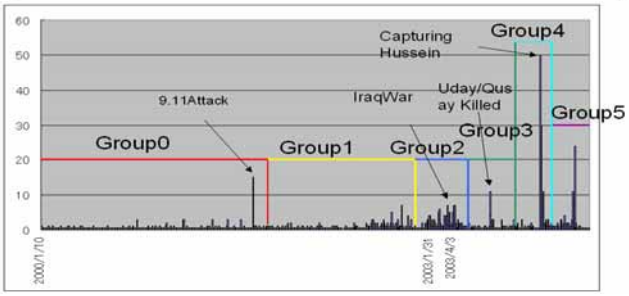

図２ Hussein のクラスタリング結果
Fig.2 Clustering Hussein pages

2001/12/15と2002/11/20の間の101ページのGroup１の文の例を示す．これら，すべてがサダム・フセインのいくつかの様相について述べている．

The U.S. Must Strike at Saddam Hussein.Bush planning to topple Hussein.Saddam Hussein to be overthrown by the opposition.How The US Armed Saddam Hussein With Chemical Weapons.Peasant-born Saddam relentlessly pursued prestige, power.For decades, Iraqi leader was both omnipresent, elusive.Hundreds Show Up For Anti-Hussein Rally.Bin Laden Linked To Saddam Hussein....

これらのクラスタに対して次のように解釈した．

(Group0: 2000/01/10 - 2001/12/18)
Attacks on World Trade Center and Pentagon
(Group1: 2001/12/28 - 2002/11/27)
About Saddam Hussein
(Group2: 2002/12/02 - 2003/05/14)
Start War
(Group3: 2003/05/19 - 2003/10/03)
Uday and Qusay were killed in a battle with U.S.
(Group4: 2003/10/08 - 2004/01/22)
Saddam Hussein captured
(Group5: 2004/01/26 - 2004/03/22)
After Getting Hussein

これらの解釈は非常にクラスタに対応したものであると言える．実際，図2で示されるように，特有の問題は適切なクラスタで発生している．

5.3 クラスタの解釈

次に，上述クラスタをKeyGraph手法を用いて重要語を抽出し，その差分を求める．

クラスタ	0	1	2	3	4	5
重要語数	31	50	54	40	63	34

次はクラスタ１の(クラスタ０との)差分である．

weapon,militari,iran,2002,inspector,intern,bush,document,famili,russian,nuclear,washington,threat, 2003,offici,16,kamel,christiansciencmoni,tore,claim,control,defect,march,missil,opposit,terrorist,plan,terror,senat,agreement

これらはステミング直後の状態であり，そのままでは理解しにくいが，ここで得られた重要語集合は，辞書や背景知識などを用いて抽象化・集約化されて統合できる．ここでは，これを次のように人手で要約する:

武装:weapon,military,plan
国際:russian, iran, internaltional,
アメリカ国内:senat,bush,tore,claim,control.Defect, washington
UnitedNations:document,inspector,agreement,terro, terrorist,nuclear,missile,opposite,threat
報道:ChristianScienceMonitor
イラク:famili,kamel

クラスタ１はブッシュのテロ支援国家，ならず者国家発言があった時期であり，“terrorist””nuclear missile”が現れる．これは，先に示した人間による解釈(About Saddam Hussein)を相当程度精密に記述したものである．同様に，クラスタ２(Start War)はテロリストとの関連,大量破壊兵器疑惑，大規模戦闘の開始が話題になった時期であり武装に関する語や，”Bin-Laden””WeaponMassDestruction”が現れる．

武装:enemy,capture,attempt,army,defense,aggressive
国際:world
アメリカ国内:leader,nation
イラク国内:author,coalit,Kurd,Bin-Laden,Amicu, Dictator party
報道:report,talk,live,fact
UnitedNations:WeaponMassDestruction, Answer

クラスタ３（Uday and …)はウダイとクサイの死亡した時期であり２人に関連した語が現れる．

武装:recruit,military,oper,troop
ウダイとクサイ:July, Husseins, son, udai,qusai
イラク体制:bremer,power,intelligence,intelligentserv, mukhabarat,secure,

クラスタ４(Saddam Captured)はフセインが拘束された時期であり報道分野の語，イラクの体制に関する語，フセイン拘束の状況を示す語が現れる．

武装:soldier,attempt
国際:arab,world,countries,intern
アメリカ国内:bush,polit,polici
UnitedNations:weapon, document
報道:video,article,report,copyright, ChristianScienceMonitor,site, work
フセイン:captur,family,sunday,death,trial,hole, crime,tikrit
イラク体制:administr,govern,leader,nation,coalit,regim

クラスタ５(After getting Saddam)は大量破壊兵器の証拠，イラク復興が話題になった時期であり，それらの話題をとらえた語が現れる．

往来:visit,com
支援・体制:redcross,author,ICRC
UnitedNations:ICRC,evid

これらの結果から，得られた語は予め与えた解釈を精緻に述べるものであり，直感的に捕らえやすいものとなっている．すなわち，提案手法の有効性が示せたと言える．

6. 結論

本稿では，検索語を与え検索エンジンからWebページ集合を取得し，時制Webページ集合から事象の抽出を行う方法と，KeyGraphを用いてクラスタの解釈を行う方法を提案した．

最初に，Webページの有効時間の推定を行い，予備実験により経験的な推定方法P_{CU} を採用した．次に，K-meansアルゴリズムによりクラスタ化しKeyGraphに基づいてそれらの解釈を行った．実験に基づく結果は，本手法が有効であることを示し，時制Webページから正確で適切に事象を抽出できることを意味している．クラスタ解釈で得られた重要語集合の抽象化・集約化ができるならば，その可読性は一段と改善できるであろうと予測することができる．

[文献]

[1] Popescul, A, Ungar, L.H.: Automatic Labeling of Document Clusters,unpublished
[2] Allan, J., Carbonell,J., Doddington, G., Yamron,J. and Yang,Y.: Topic Detection and Tracking Pilot Study: Final Report, proc.DARPA Broadcast News Transcription and Understanding Workshop (1998)
[3] Grossman,D. and Frieder,O.: Information Retrieval—Algorithms and Heuristics, Kluwer Academic Press, 1998
[4] Jain, A.K., Murty, M.N. et al.: Data Clustering, ACM Comp.Surveys 31-3, 1999, pp.264 - 323
[5] Kleinberg, J.M. : Authoritative Sources in a Hyperlinked Environment, JACM 46-5, 1999
[6] Mani, I.: Automatic Summarization, John Benjamins, 2001
[7]大沢幸生、他: KeyGraph - 語の共起グラフの分割統合によるキーワード検出、電子情報通信学会論文誌D-I、J82-D-I2,pp.391-400,1999
[8] NIST (National Institute of Standrads and Technology): www.nist.gov/speech/tests/tdt/
[9] Radev, D. and Fan, W. : Automatic summarization of search engine hit lists, proc ACL'2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval, 2000, Hong Kong
[10] Yang, Y., Pierce, T. and Carbonell,J.: A Study on Retrospective and On-Line Event Detection, proc. SIGIR-98, ACM Intn'l Conf. on Research and Development in Information Retrieval, 1998

森 正輝 Masaki MORI
法政大学工学研究科電気工学専攻修士課程在学中．セマンティック Web の研究に従事．日本データベース学会学生会員

三浦 孝夫 Takao MIURA
京都大学理学部，工学博士(東京大学)．現在，法政大学工学部情報電気電子工学科教授．データモデル，知識表現，演繹データベース，複合オブジェクトなどの分野の研究に従事．電子情報通信学会，ACM 各会員．著書に"データモデルとデータベース"(全 2 巻，サイエンス社)

塩谷 勇 Isamu SHIOYA
東京電機大学大学院修士課程了．現在，産能大学経営情報学部教授．時系列モデルの同定，論理プログラミング，グラフ文法，論理データベースの研究に従事．電子情報通信学会，ACM 各会員.