[Full Paper] 時制クラスタのトピック追跡

森　正輝† 三浦 孝夫† 塩谷 勇††

† 法政大学 工学研究科 電気工学専攻 〒 184–8584 東京都小金井市梶野町 3-7-2
†† 産能大学 経営情報学部 〒 259–1197 神奈川県伊勢原市上粕屋 1573
E-mail: †{i04r3246,miurat}@k.hosei.ac.jp, ††shioya@mi.sanno.ac.jp

あらまし 現在、トピック追跡の手法として文書分類の手法を利用した方法、また、クラスタに対してのラベル付けの方法として、単語の出現頻度や単語の共起性を考慮したラベル付けなどの手法が提案されている。本稿では、Web ページから時制クラスタを生成し、各クラスタで土台となる語集合を抽出しトピックの追跡を行い、KeyGraph と Suffix Tree、単語の並び、時間軸を考慮することで Web ページ集合に対して抽象度の高いラベル付けを行う。本稿では、実験により提案手法の有用性を示す。

キーワード Web マイニング, TDT, ラベリング

[Full Paper]Topic Tracking from Temporal Clusters

Masaki MORI†, Takao MIURA†, and Isamu SIOYA††

† Dept.of Elect.& Elect. Engr., HOSEI University 3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
†† Department of Management and Information Science, SANNO University 1573, Kamikasuya, Isehara city, Kanagawa 259–1197 Japan
E-mail: †{i04r3246,miurat}@k.hosei.ac.jp, ††shioya@mi.sanno.ac.jp

Abstract In this investigation, we propose a new approach of topic tracking to extract and summarize and track events from a collection of Web Pages. Given a set of Web pages, there have been several methods for topic tracking proposed so far. Here we examine Web pages and obtain *valid* timestamp, and detect events by means of clustering. Next, we discuss a novel technique to track events by using KeyGraph based on the clusters. Then we abstract clusters by using both KeyGraph and SuffixTree based on the clusters. We show some experimental results.

Key words Web Mining, TDT, Abstracting

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、日を追うごとに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る。Web ページ集合の表している内容について、いつ何が起こったのかを利用者が知っている場合も知らない場合も、利用者の求める Web ページ集合を見つけ出すことは非常に労力を必要とする。このため Web ページ集合の内容を素早く容易に把握する研究が近年注目を浴びている [2], [10], [13]。

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に関する Web ページの URL 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のを手助けするために、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。この巨大なデータベースを用いた検索により情報重複の問題を軽減させることができる。しかしながら、新たに非常に長い検索結果のリストを出力してしまうという問題が発生する。利用者は、得られた検索結果をブラウズし有益な Web ページを探すのだが、多くの場合、途中で断念してしまう。実際、ほとんどの場合利用者は、最初の 10 又は 20 ページだけをブラウズして有益な Web ページを探し出すと言われており、この問題は深刻である。言い換えると、ページのランキングだけで選択が決定されており、この決定方法が重要な問題となっている。現在では、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手法などいくつかの手法が提案されている [5]。

しかし、これらの手法はトピックを把握するのに適した手法ではない。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 Web ページをクラスタに分類し、クラスタを追跡することでトピック追跡ができ、さらにクラスタの情報を要約できたならば、利用者が、検索結果をより効果的に容易に吟味することができ、負担も軽減されると考えられる。

更に、ページの有効時間を類推することができれば、内容を時間に沿って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], [9]。TDT 研究プロジェクトでは、時間軸上で自動的にニュースストリームからトピックの意味の構造を抽出することを目的とした議論がされている。

我々は、これまでに検索エンジンから得られた検索結果から時制クラスタを抽出し、KeyGraph と Suffix Tree に基づく手法を用い各クラスタから主張語を抽出しクラスタの自動解釈を行う手法を提案している [7], [8]。本稿では、時制的な側面を持つ Web ページ集合からトピック追跡を行う手法を提案し、SuffixTree を用い単語の並びを用い、主張語を考慮した抽象度の高いラベル付け（要約あるいは抽象化）を行う。

本稿では、2 章でトピック追跡とラベル付けの意義と目的、3 章で時制クラスタの抽出、4 章でトピック追跡、5 章でラベルの決定、6 章で実験と考察を行い、7 章で結論とする。

2. トピック追跡とラベル付けの意義と目的

2.1 考 え 方

本稿では、時制クラスタに対して Web ページの基本概念となる単語を考慮したトピック追跡手法を提案しクラスタに対してラベル付けを行う。

トピック追跡手法として、分類法を用いたトピック追跡手法が提案されている。この手法は訓練データからトピックの特徴を獲得し、未知のデータに対して同一トピックかどうか判断しトピック追跡を行うものである。この手法を用いた場合、初期値に依存するため時間の経過によるトピックの基本的な概念の変化には対応できない。

Web ページ集合にトピックが 1 つだけの場合、時間軸でクラスタリングし得られた全ての事象が 1 つのトピックに対応するので容易にトピック追跡が行える。しかし、本稿では Web ページ集合を検索エンジンに検索語を与えて収集するため、収集された Web ページ集合は複数のトピックから構成される Web ページ集合ある。そのため、時間軸でクラスタリングし得られた事象がどのトピックに関するかを判断しなければトピックを追跡できない。更に、事象は時間の経過に従って、”関連の薄かった事象が合併””1 つの事象が分離””新たなトピックの事象が発生”” トピックが消滅”する。したがって、複数のトピックが混在する Web ページ集合からトピック追跡を行うには、古い概念を捨てながら新しい概念を取り入れることが非常に重要となる。

本稿では、古い概念を捨て新しい概念を取り入れ、複数のトピックが混在する Web ページ集合から正確なトピック追跡を行う。これにより、利用者は複数のトピックが混在する検索結果からトピックごとに事象を分けることができ、同一トピックの事象を時間順に辿ることでトピックの流れを把握することができるため、有益な Web ページを見つけやすくなる。

更に、事象に対してラベルを付けることできれば、利用者は更に事象の内容、トピックの流れを把握できるようになり更に利用者の手間は軽減される。

ラベル付け手法として、Web ページ中で発生頻度の高い語をラベルとする方法が考えられる。しかし、発生頻度の高い語だけで Web ページの内容の詳細を示すことは難しい。検索エンジンに検索語を与えて得られる Web ページは非常に類似性が高く、各 Web ページ集合で発生頻度の高い語にほとんど差異はない [7]。したがって、語の発生頻度だけでラベル付けを行うのは適した方法ではない。Web ページの主張を捕らえた単語を抽出することができれば、利用者の手間も軽減されると考えられる。

更に、利用者に Web ページ集合の意味を容易に把握するには、単語だけのラベルよりも、単語の並びで意図を表現したラベルの方がよいことが知られている [12]。

本稿の基本的なアイデアは 3 段階からなる。まず検索エンジンに検索語を与え、得られた Web ページの有効時間を推定し、時間軸でクラスタリングを行い時制クラスタを得る。これは事象に対応しやすいことに注目すべきである。次に、各クラスタに対して、その基本概念を捕らえた単語を KeyGraph で抽出しサブクラスタを構築し、隣接する時制クラスタのサブクラスタ同士の関連性を発見することでトピック追跡を行う。最後に、サブクラスタの主張を捕らえた語を KeyGraph で抽出し、単語の並びを Suffix Tree を用いることで抽出しラベル付けを行う。

2.2 準 備

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 [11]。

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(注1)。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主張語と呼ぶ(注2)。更に、クラスタ化された基本概念と主張語の共起度を計算し、共起リンクに値を与え共起リンクの和をとる。最終的に、共起リンクの和の上位語を土台と主張を結びつける重要語(注3)とする。なお、本稿では同一トピックを追跡する為に基本概念に注目し、クラスタの主張を捕らえる為に主張語に注目する。

［例題 1］ 以下に示す 3 つの文書に対して KeyGraph を生成する。

文書 1: human ate carrot.

文書 2: rabbit ate carrot too.

(注1)：KeyGraph では「土台」と呼ぶ。

(注2)：KeyGraph では「屋根」と呼ぶ。

(注3)：KeyGraph では「柱」と呼ぶ。

文書 3:　human ate rabbit too.

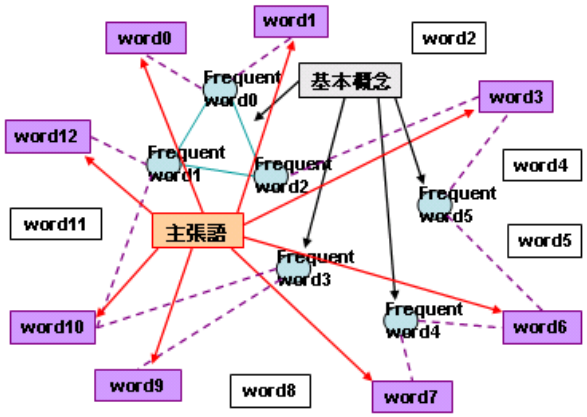

図 1　基本概念と主張語

文書から不要語除去、ステミングを行った後、単語単位で KeyGraph を形成する。ステミングとは、単語の語幹だけを残すことである。例えば、”swims””swimming””swimmer”などの単語は語幹だけが残り”swim”となる。3 回以上出現する語を基本概念とし、主張語の抽出を行う。図 2 に例題の KeyGraph を示す。

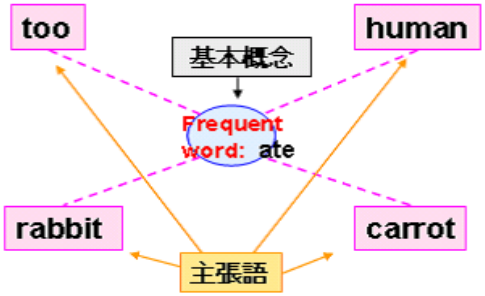

図 2　基本概念と主張語

KeyGraph に基づき、基本概念「ate」主張語「carrot」「human」「rabbit」「too」重要語「ate」が得られる。

Suffix Tree(接尾辞木) とは、文書に出現する単語をノードとし全ての単語の並びを表した Tree であり、文字列 S の Suffix Tree とは全ての S の接尾辞を含む木である。この木はルートから始まる方向性を持ち、 中間ノードは少なくとも 2 つ以上の子供を持ち、全ての枝はラベルを持つ。ただし同じノードから同じ言葉で始まる枝は無い。また S の接尾辞 s に対応するラベル s の接尾辞ノードを持つ。

［例題 2］　以下に示す 3 つの文書に対して Suffix Tree を構築する。

文書 1:　human ate carrot.
文書 2:　rabbit ate carrot too.
文書 3:　human ate rabbit too.

文書から不要語除去、ステミングを行った後、単語単位で Suffix Tree を形成する。

各ノードは、それぞれ固有の単語の並びを持つ。以下に、複数の文書で構成されるノードの詳細を示す (表 1)。

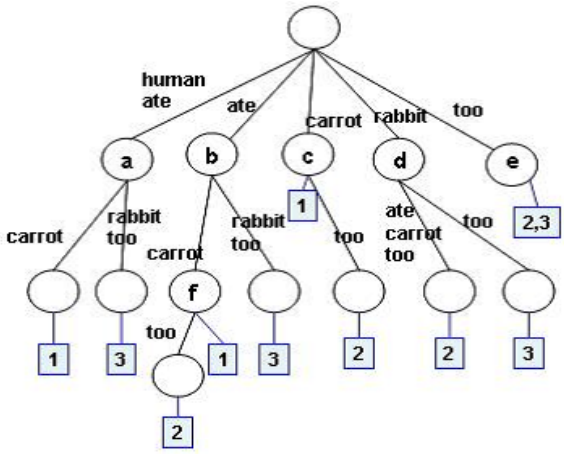

図 3　Suffix Tree

ノード	単語の並び	文書
a	human ate	1,3
b	ate	1,2,3
c	carrot	1,2
d	rabbit	2,3
e	too	2,3
f	ate carrot	1,2

表 1　各ノードの詳細

3.　時制クラスタの抽出

本稿で論じる時制クラスタとは、トピックに関する文書を時間軸でクラスタ化したものである。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ことはよく知られている [2]。すなわち、事象はしばしば時制クラスタに対応する。我々は既に、検索エンジンに検索語を与えて得られる検索結果から時制クラスタを抽出する手法を提案している [7]。

まず Web ページの有効時間の推定を行う。全ての Web ページを解析し内容時間を抽出、内容時間を抽出できなければ URL より作成時間を抽出し有効時間とする。内容時間も作成時間も抽出できない Web ページは除去する。内容時間とは Web ページの内容が意味する時間であり、それぞれの文章の最初に明示的に出現しているタイムスタンプである。作成時間は Web ページが作成された時間であり、経験的に URL に作成時間が一部として現れる。次に、時間軸上で K-means 法を用いてクラスタリングを行う。この時、構成要素の少ないクラスタを無視する。

この手法の有効性はすでに実験により確かめており、時制クラスタがうまく生成できることを確認している [7]。しかし、さらに本稿では、提案手法の評価のために、残ったクラスタのラベルを人手で与えるものとする。検索語を含む文章を抽出し、人手でラベルを決定する。人手によるラベルの評価は実際の事象が適切なクラスタに対応しているかで評価する。

4. トピック追跡

4.1 基本概念の抽出

KeyGraph に基づき基本概念の抽出を行う。文書 D から不要語処理・HTML タグ除去・ステミング処理を行った後、得られた語集合 W から、上位定数個の頻出単語 $w_1, .., w_N$ を抽出してその共起度を計算する。すなわち、文 (sentence) s ごとに語 w_i, w_j の出現回数 $|w_i|_s, |w_j|_s$ を求め、次の共起度 $co(w_1, w_j)$ を得る。

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

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。この定義からわかるように、各土台とは頻出語で共起度でクラスタ化した語集合であり、よく知られた概念の集合体（基礎概念）に対応するとみなすことができる。

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したい。

4.2 サブクラスタの構築

各時制クラスタ内で基本概念を用いてクラスタリングを行いサブクラスタを構築する。クラスタリング手法として Complete-Link クラスタリングを用いる。Complete-Link クラスタリングとは、2 つのクラスタの要素で最も類似していない要素同士が閾値を超えていれば 2 つのクラスタを合併する（図 4)。本稿では基本概念の類似度をコサイン値で算出し、閾値を経験的に 0.1 以下とする。 クラスタリングを行った後、基本概念を抽出

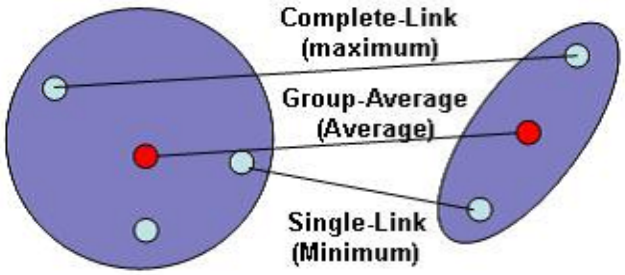

図 4 Complete-Link クラスタリング

できたページの総数の 10 %以上のページ数のサブクラスタを抽出する。

4.3 トピック追跡

トピック追跡を隣接する時制クラスタのサブクラスタを比較し類似するサブクラスタを追跡することで行う。隣接する時制クラスタのサブクラスタを比較することで、古い概念を捨てトピック追跡を行えるようになり、トピックの発生、消滅、2 つの事象の合併、1 つの事象の分離を考慮したトピック追跡表現が可能となる。隣接する時制クラスタのサブクラスタの代表点をコサイン値で比較する。経験的に 0.57 以上の値であれば 2 つのサブクラスタは同一トピックについて論じており関連性がある。

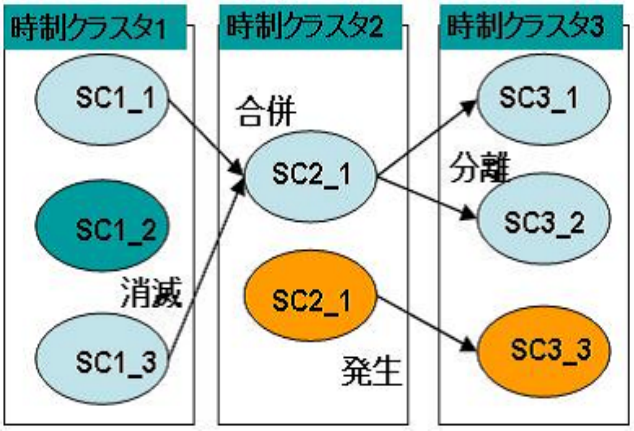

図 5 Topic の消滅、発生、合併、分離

5. ラベルの決定

5.1 主張語の抽出

KeyGraph に基づく手法で、基本概念（土台 g ）を抽出した後、KeyGraph に基づき主張語を抽出する。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したい。$|w|_s$ 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　$w \in g$ 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であるとする。すなわち、

$$key(w) = probability(w | \cap_{g \subset G} g)$$

つまり

$$key(w) = 1 - \Pi_{g \subset G}(1 - \frac{based(w, g)}{neighbor(g)})$$

ここで $\frac{based(w,g)}{neighbor(g)}$ は土台 g を用いるときに語 w も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主張語とみなす。本稿では、各 Web ページを文とみなし、KeyGraph によりサブクラスタから抽出した語の全てを主張語とする。

5.2 単語の並びの抽出

Suffix Tree を用いて単語の並びを抽出する。Web ページから不要語、HTML タグを取り除きステミングを行った後、単語単位で Suffix Tree を形成する。

本稿では、各サブクラスタごとに単語の並びが 6 単語までを対象とし Suffix Tree を形成する。そして、各サブクラスタを構成する Web ページの総数 30 パーセント以上の頻度の単語の並びを抽出する。

5.3 ラベルの決定

KeyGraph に基づく主張語、Suffix Tree から得られた単語の並びをそれぞれ抽出した後に、ラベルの決定を行う。まず、Suffix Tree から得られた単語の並びに対して、主張語を考慮してスコアを次のように定義する：

$$score(p) = (|w|_p + |s|_p) \times |p|_c$$

p は Suffix Tree から得られた単語の並び、$|w|_p$ は p の単語の並びを構成する単語数、$|s|_p$ は p の中に含まれる主張語の数、$|p|_c$ はクラスタ c での p の発生回数を示す。

本稿では、スコアの高い単語の並びを用いてサブクラスタのラベル付けを行う。追跡可能なクラスタは同一トピックを論じたものであるため、追跡可能なサブクラスタは相互に類似性が高く、出現頻度だけに依存しない提案手法でも、得られた単語の並び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、「時制クラスタのラベル付け」を「短期的な概念変化の状況の記述」と考え、追跡可能なサブクラスタは、直前の追跡可能なサブクラスタにおける単語の並びの集合の差分をラベル付けに用いる。

本稿では、追跡可能なサブクラスタは直前の追跡可能サブクラスタとの差分をラベルとし、それ以外は、単語の並びの集合のスコア値の高い上位 50 %をラベルとする。この手法の有効性はすでに実験により確かめており、主張語と単語の並びを考慮したラベル付け手法が有効であることを確認している [8]。

6. 実　　験

6.1 手　　順

本稿では、提案手法の有用性を示すために Google から 1000 ページの Web ページを取得し実験的な結果を論じる。

検索エンジン Google に検索語「hussein」を与え、得られた結果より、リンク切れ、Weblog、時間情報のない Web ページを除去した後、有効時間の推定を行いクラスタリングを行う。次に、得られた時制クラスタから提案した手法でトピック追跡を行いラベル付けを行う。このときラベルの評価のために、時制クラスタのラベル付けを人手でも行う。トピック追跡の評価は得られたラベルを基に行う。

6.2 時制クラスタの生成

はじめに、時制クラスタの生成を行う。検索エンジンに検索語「hussein」を与えクラスタリングを行った結果を以下に示す。

GroupID	ページ数	内容時間	作成時間
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

また、各クラスタごとに特徴的なラベルを人手により付与する。2001/12/15 と 2002/11/20 の間の 101 ページの Group1

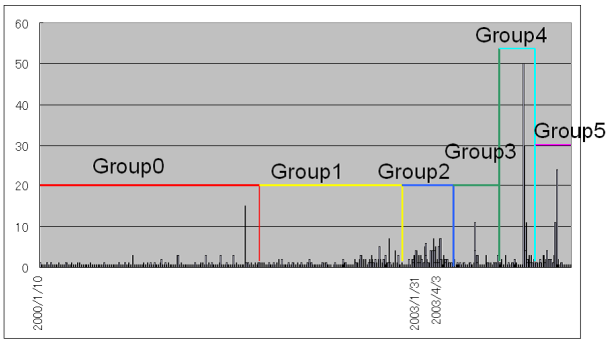

図 6　Hussein のクラスタリング結果

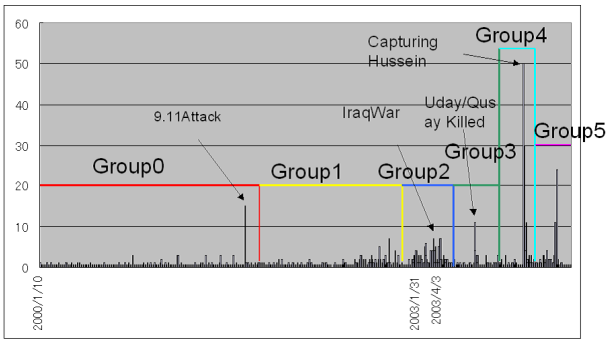

図 7　実際の事件と時制クラスタの対応

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
Opposing Saddam Hussein
[Hussein Ibish:] U.S. Arabs' Firebrand
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,
....
```

次に、以下のように全てのクラスタを解釈した。

(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 7 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。

6.3 トピック追跡

各時制クラスタごとにサブクラスタを構築し詳細を表 2 に示す。表 2 の「処理を行ったページの数」とは各時制クラスタで KeyGraph を構築できたページの数である。サブクラスタ ID は前の数が時制クラスタの ID を示し、後の数が時制クラスタ内のサブクラスタの ID を示す。つまり、「0-0」ならば時制クラスタ 0 のクラスタ 0 という意味になる。

サブクラスタ ID	処理を行ったページ数	ページ数
0-0	33	11
1-0	50	19
2-0	81	9
3-0	33	6
3-1	33	8
4-0	127	19
5-0	42	6

表 2　サブクラスタ詳細

次に、隣接する時制クラスタのサブクラスタの類似度を表 3、それに基づくトラッキングの結果を図 8 に示す。

比較するサブクラスタ	類似度
0-0 1-0	0.07
1-0 2-0	0.76
2-0 3-0	0.57
2-0 3-1	0.43
3-0 4-0	0.69
3-1 4-0	0.50
4-0 5-0	0.57

表 3　サブクラスタの類似度

図 8 よりサブクラスタ 0-0、1-0、3-1 でトピックが発生し、サブクラスタ 1-0 で発生したトピックは、サブクラスタ 2-0、3-0、4-0、5-0 と順にサブクラスタを追跡している。

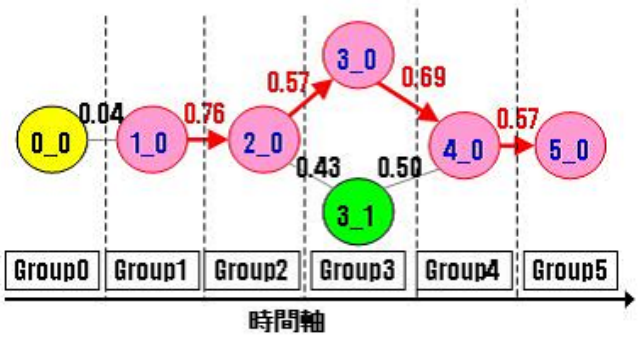

図 8　Tracking 結果

サブクラスタ ID	単語の並び	主張語	スコア値上位 50 %
0-0	184	160	91
1-0	240	218	125
2-0	243	68	131
3-0	148	87	75
3-1	346	40	213
4-0	110	121	57
5-0	306	19	81

表 4　抽出された語

6.4 ラベル付け

Suffix Tree に基づいてサブクラスタの 30 %以上のページで出現する単語の並び、KeyGraph に基づく主張語全てを抽出しスコアの高い上位 50 %の単語の並びを抽出する (表 4)。

次に追跡可能なサブクラスタについてサブクラスタの差分を抽出する。(表 5)。

サブクラスタ ID	ラベル
2-0	58
3-0	44
4-0	27
5-0	79

表 5　ラ　ベ　ル

6.5 実 験 結 果

次はサブクラスタ 2-0 の (サブクラスタ 1-0 との) 差分である。

live ,stop ,forc unit ,iraqi peopl ,presid saddam , 12 year ,arm forc ,iraq lead ,question comment articl ,war end ,hour ,coalit ,comment ,show ,ambassador , captur ,comment articl ,econom sanction ,freedom ,friend colleagu ,kurd ,question comment , simpler version ,12 ,2003 ,compani ,conflict ,continu ,dai ,dictat ,econom ,final ,friend , histori ,long ,march ,million ,question ,town ,washington ,arm ,articl ,di ,fight ,found , free ,intellig ,iranian ,kill ,pass ,power ,prove ,refus ,remain ,resolut ,start ,thousand ,todai

これらはステミングされた状態であるので、そのままでは理解しにくいが、さらに得られ単語の並びの集合は、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(注4)。本稿ではこれらの単語の並びを人手で要約する。

最初に、サブクラスタ 1-0、2-0、3-0、4-0、5-0、の実験結果について議論する（図 8)。

サブクラスタ 1-0

武装:　forc, militari, armi, troop, attack

国際:　arab, chairman, world, univers, kuwait, prime minist

(注4)：たとえば Wordnet などの辞書を活用すればよい。
http://wordnet.princeton.edu

アメリカ: unit state, presid bush, claim, american, minist, secretari, launch
イラク: saddam hussein, gulf war, iraqi, iraqi leader, regim, iraqi govern, baghdad
UnitedNations: chemic biolog weapon, weapon mass destruct, secur council, weapon inspector, unit nation, inspect, nuclear

これらは、サブクラスタ 1-0 で得られた単語の並びを人手で要約したものである。この時期は、イラク戦争の開戦前の時期であり、”chemic biolog weapon””weapon mass destruct””weapon inspector” などのラベルが得られている。よって、サブクラスタ 1-0 は「イラク戦争直前」について論じている。

サブクラスタ 2-0

武装: forc unit, arm forc, coalit, power, start,
国際: iraniraq,friend colleagu, stop, prove, conflict,
アメリカ: free, freedom, washington, ambassador, question, intellig,
報道: live, comment articl,
イラク: dictat, presid saddam, iraqi peopl, kurd, refus,

サブクラスタ 2-0 は、大規模戦闘が始まった時期である。”freedom””presid saddam””start”などのラベルから、サブクラスタ 2-0 は「イラク戦争開始」を表すサブクラスタである。

サブクラスタ 1-0 とサブクラスタ 2-0 の内容を比較すると、「イラク戦争直前」を表した内容であるサブクラスタ 1-0 と、「イラク戦争開始」を表すサブクラスタ 2-0 は、関係性があり同じトピックについて論じられていると判断できる。

サブクラスタ 3-0

武装: command, oper, combat, troop iraq, 4th infantri, infantri divis
アメリカ: report iraq,
フセイン: son udai qusai, captur kill, iraqi leader, saddam husseins, saddams
イラク: baathist, tikrit, baath parti, mosul, secretari

サブクラスタ 3-0 は、ウダイとクサイが死亡した時期である。サブクラスタ 3-0 では、ウダイとクサイを表した”son udai qusai””saddam husseins”がラベルとして得られているが、これらは、”サダムフセインの息子達”と言い表した表現と言え「サダムフセイン」と、多数の武装に関するラベルから「イラク戦争」について論じている判断できる。

サブクラスタ 2-0 とサブクラスタ 3-0 の内容を比較すると、「イラク戦争開始」を表すサブクラスタ 2-0 の内容と、「イラク戦争とサダムフセイン」について論じているサブクラスタ 3-0 は、関連性があり同じトピックであるとラベルから判断できる。

サブクラスタ 4-0

武装: soldier, coalit forc
国際: nation, unit
アメリカ: bush, paul bremer
フセイン: saddam hussein captur, captur saddam hussein, hide, forc captur, captur saturdai, arrest, captur forc,
イラク: dictat, iraqi peopl

サブクラスタ 4-0 は、サダムフセインが拘束された時期である。サブクラスタからは、”saddam hussein captur””hide””arrest” など、サダムフセインが拘束されたことを表したラベルが得られていることから、サブクラスタ 4-0 はサダムフセイン拘束について論じられているサブクラスタである。

「サダムフセインとイラク戦争」と「サダムフセイン拘束」を表すサブクラスタ 3-0 と 4-0 を比較すると、2 つのサブクラスタとも、サダムフセインとイラク戦争について関連があり 2 つのサブクラスタは同じトピックであると言える。

サブクラスタ 5-0

武装: pow, 10000 prison, occup saddam hussein collabor 12, prison war,
UnitedNations: red cross visit saddam hussein, icrc visit,
フセイン: wrote letter famili, good health, iraqi leader saddam hussein, health condit,
イラク: oust iraqi leader, shape futur iraq,

サブクラスタ 5-0 では、”good health””red cross visit saddam hussein””health condit”など、拘束された後のサダムフセイン様子を表現したラベルが得られている。よって、サブクラスタ 5-0 は「サダムフセイン拘束後」を表したサブクラスタであると言える。

これらから、サブクラスタ 1-0、2-0、3-0、4-0、5-0 は、戦争が始まってフセインが拘束されたその後までの、イラク戦争の一連の流れを表しているので、提案手法によりトピック追跡が出来たと言える。

サブクラスタ 0-0

武装: enemi, weapon, militari, power
国際: middl east, arab, israel,
アメリカ国内: washington, unit state, unit nation, claim, secur, presid, prime minist
イラク: saddam hussein, baghdad, baath parti, oil, invas kuwait, econom sanction, iraqi leader, kurd, gulf war
9/11: trade, world, peopl, attack, forc,

サブクラスタ 0-0 は、「アメリカとイラクの関係」を表す”econom sanction””gulf war”や”attack””world””forc”などの「同時多発テロ」を表す単語が現れている。

しかし、同時多発テロから直接イラク戦争に直接繋がった事実はなく、「同時多発テロとアメリカとイラク関係」を表すサブクラスタ 0-0 から、「イラク戦争直前」を表すサブクラスタ

1-0 を追跡できないのは妥当であり、サブクラスタ 0-0 は別のトピックと言える。

サブクラスタ 3-1

武装: forc, arm, troop, attack, secur forc, arm, chief, militari, troop, soldier, weapon, fight, command, oper
国際: islam, nation, unit, world, state, arab,
ウダイとクサイ: udai qusai hussein, death, kill, juli, qusai hussein, brother, prospect, suspect
アメリカ: intellig servic, bush, secur servic, intellig servic, american, claim , presid
イラク: baath parti, baghdad, iraqi peopl, iraqi intellig, govern, civilian, defens
報道: author, inform, report,

サブクラスタ 3-1 は、ウダイとクサイが死亡した時期であり、"udai qusai hussein""brother""qusai hussein""kill""juli" などのウダイとクサイが死亡したことを言い表したラベルが得られていることから「ウダイとクサイ死亡」について論じたサブクラスタであると言える。

イラク戦争に関した内容のサブクラスタであるが、イラク戦争直前からフセイン拘束後までの一連の話の流れに必ずしも必要な内容ではなく、トピック追跡を行わないのは妥当あり、別のトピックであると言える。

6.6 評　　価

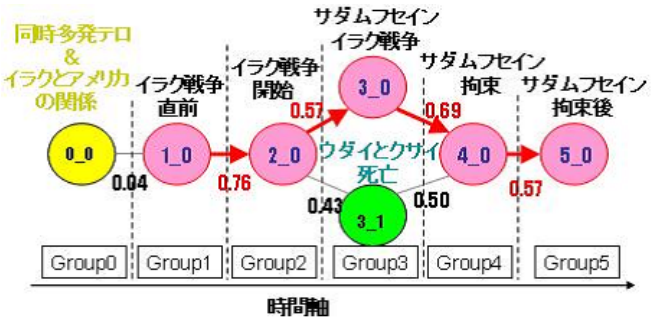

図 9　Topic の結果

実験結果より、3 つのトピックが得られ（図 9）、隣接するクラスタのサブクラスタとの関係性、他のトピックとの関係性を検証し 3 つのトピックが、固有のトピックについて論じていることを示した。これらから判断し、提案した手法により複数のトピックの混在する Web ページ集合からトピック追跡が可能になったと言える。すなわち、提案した手法でトピック追跡が自動的にできれば、利用者は、トピックの追跡が容易に行え、検索で得られた Web ページを簡単に把握できるようになる。

本手法が想定する主要な前提は、「時制クラスタは事象に対応する」という点にある。複数のトピックを含む Web ページ集合 (「リンカーン」は自動車、人物の双方を含む)、あるいは時制的側面の弱いトピック (「ロサンゼルス」だけではメジャーリーグ以外に時制的な扱いができない) に対しては、適応外であろう。KeyGraph あるいは SuffixTree に基づく本手法が、広範囲な対象に対して的確に機能するためには、事象抽出手法との連動が必要となろう。

7. 結　　論

本稿では、検索語を与え検索エンジンの結果を時制クラスタを取得し、各クラスタごとに KeyGraph に基づく手法で抽出した基本概念を用いてサブクラスタを構築し、古い概念を捨て、新しい概念を取り入れて複数のトピックの混在する Web ページ集合から、トピック追跡を行う手法を提案し Suffix Tree を用いて単語の並びを抽出し、事象に対して KeyGraph に基づく手法で抽出した主張語を考慮したラベル付けを行った。

最初に、各クラスタで KeyGraph に基づいて抽出された土台語を用い、Complete-Link クラスタリングを行い、サブクラスタを構築した。次に、直前の時制クラスタのサブクラスタとの類似度を比較することで、古い概念を捨て、新しい概念を取り入れ、トピックの追跡を行った。実験に基づく結果は、提案した手法が有効であることを示し、時制クラスタのトピック追跡が可能であることを意味している。

謝　　辞

本研究の一部は文部科学省科学研究費補助金 (課題番号 16500070) の支援をいただいた。

文　　献

[1] Alexandrin Popescul,Lyle H. Ungar.: 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] 森 正輝, 三浦 孝夫, 塩谷 勇: Web ページからの時制クラスタの解釈, 日本データーベース学会 Letters Vol.3, No.2, pp.109-112,2004
[8] Masaki Mori, Takao Miura, Isamu Shioya: Abstracting Temporal Clusters, ITA, 2005
[9] NIST (National Institute of Standrads and Technology): `www.nist.gov/speech/tests/tdt/`
[10] 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
[11] 大沢幸生 : KeyGraph　一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
[12] Oren Zamir and Oren Etzioni.: Web Document Clustering: A Feasibility Demonstration, SIGIR 1998: 46-54
[13] 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