Suffix Tree Clustering を用いた Web ページ集合のラベル付け

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、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さまざまな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、KeyGraph および Suffix Tree Clustering を行う為の手順により Web ページ集合に対して抽象度の高いラベル付けを行う。そして、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

Abstracting Web pages by using KeyGraph and SuffixTree Clustering

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 Given a set of Web pages, there have been several methods for labeling proposed so far. In this inevestigation, we discuss a novel technique to summarize and abstract the pages by using both KeyGraph and Suffix Tree Clustering based on temporal clustering. 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], [9], [12]。

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得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], [8]。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 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]。本稿では、時制的な側面を持つクラスタに Suffix Tree Clustering (STC) を行う為の手順を用い、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) 手法を提案する。

本稿では、2 章でラベル付けの意義と目的、3 章で時制クラスタの抽出、4 章でラベルの決定、5 章で実験と考察を行い、6 章で結論とする。の決定方法を論じる。

2. 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 ページの主張と単語の並びを考慮したラベル付け行う手法を提案する。

ラベルの無いクラスタから、利用者が有益な Web ページを見つける場合、利用者が各クラスタの Web ページの内容をブラウズして確認するしかなく、非常に手間のかかる作業である。各クラスタの内容が、高度に抽象化されたラベルで表されれば、利用者が有益な Web ページを見つけやすくなる。

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、各クラスタ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 [11]。

本稿の基本的なアイデアは 2 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、その主張を捕らえた語を KeyGraph で抽出し、単語の並びを STC を行う為の手順に基づき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 [10]。

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、本稿ではクラスタの主張を捕らえるという立場から、主張語に注目する。

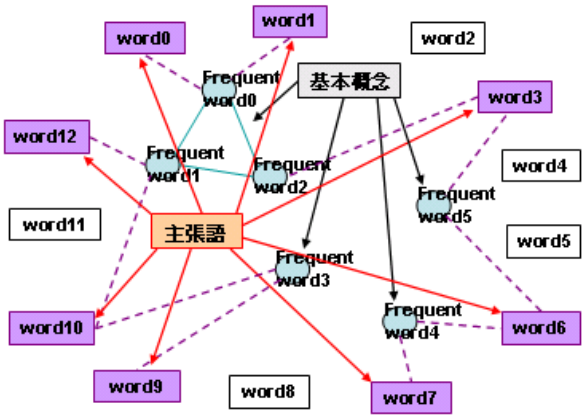

図 1 基本概念と主張語

［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.
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 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。

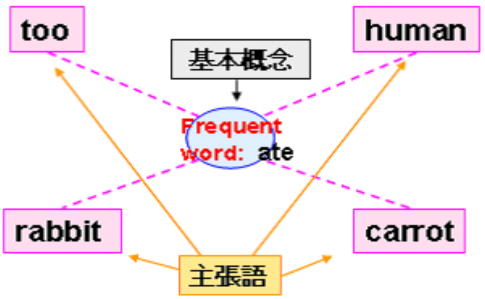

図 2 基本概念と主張語

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。

(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。

Suffix Tree Clustering (STC) とは、文書から単語単位で Suffix Tree (接尾辞木) を作りノードをクラスタリングを行う手法である [11]。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。

本稿では、単語の並びの抽出を STC を行う為の手順に基づいて行う。

［例題 2］ 以下に示す 3 つの文書の Suffix Tree を形成する。

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。

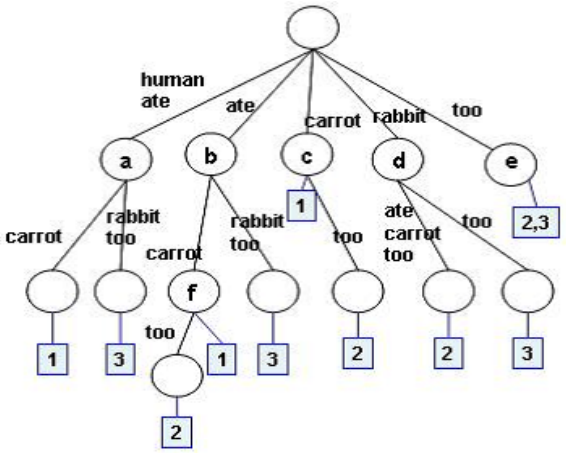

図 3 Suffix Tree

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。

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 [4]。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. ラベルの決定

4.1 主張語の抽出

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。

$|w|_s$ 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 により時制クラスタから抽出した上位 9 パーセントの語を主張語とする。

4.2 単語の並びの抽出

STC を行う為の手順に基づき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、各時性クラスタごとに単語の並びが 5 単語までを対象とし Suffix Tree を形成する。そして、各時性クラスタを構成する Web ページの総数 10 パーセント以上の頻度の単語の並びを抽出する。

4.3 ラベルの決定

KeyGraph に基づく主張語、STC を行う為の手順に基づく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、STC を行う為の手順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 STC を行う為の手順に基づいて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時制クラスタのラベル付けを行う。実験で用いる 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、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、直前の時制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、単語の並びの集合のスコア値の高い上位 9 %を差分対象とする。

5. 実　　験

5.1 手　　順

本稿では、提案手法の有用性を示すために、Google によって得られる 1000 ページの Web ページについて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。得られた時制クラスタに提案した手法でラベル付けを行う。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い、人手によるラベル、主張語だけを用いたラベル、提案手法によるラベルを比較し考察を行う。

5.2 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

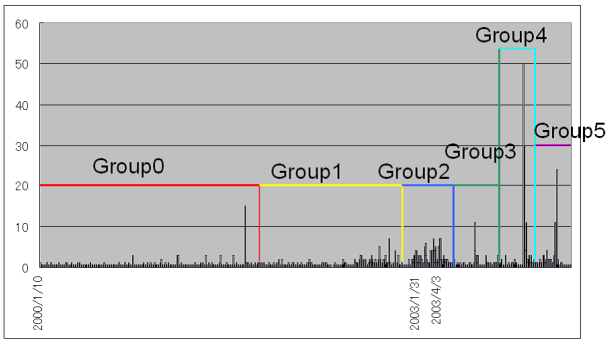

図 4　Hussein のクラスタリング結果

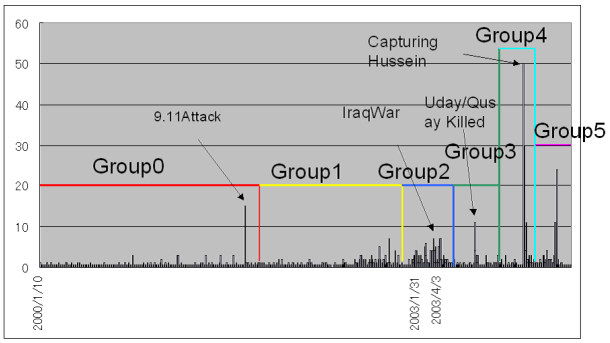

図 5　実際の事件と時制クラスタの対応

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 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 5 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 ラベル付け

はじめに、Suffix Tree に基づいて時性クラスタの 10 %以上のページで出現する単語の並び、KeyGraph に基づく主張語上位 9 %を抽出しスコアの高い上位 9 %の単語の並びを抽出する (表 2)。

GroupID	単語の並び	主張語	スコア値上位 9 %
0	1345	31	120
1	1286	50	115
2	1365	54	125
3	1284	40	130
4	1089	63	98
5	948	34	96

表 2　抽出された語

次にスコア値の上位 9 %の単語の並びの集合と主張語の差分をそれぞれ抽出する (表 3)。

GroupID	主張語差分	提案手法
1	30	50
2	27	46
3	24	65
4	41	45
5	7	41

表 3　ラ　ベ　ル

5.4 実 験 結 果

次はクラスタ 1 の (クラスタ 0 との) 差分である。

intern, militari, bush, gener abdul qassim kassem, hussein kamel, threat, offici, chemic weapon, washington, terror, pa, res, mi, inte, int, pro, le, iran, sp, ch, ca, plan, weapon inspect, saddam husseins, weapon inspector, nuclear weapon, na, gener abdul qassim, biolog weapon, stat, ho, forc captur saddam hussein, fi, cl, si, nuclear, gr, rep, bi, march, echasten feith return pentagon, chemic biolog weapon, militari action, sec, secur council, inspector, gov, terrorist, missil, milit,

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)。ここでは、これを次のように人手で要約する:

クラスタ 1 : 提案手法の結果

武装:　military, plan, military action
国際:　iran,
アメリカ:　state, bush, washington
UnitedNations:　weapon inspect, weapon inspector, nuclear weapon, chemical biological weapon, chemical weapon, threat, secur council
イラク:　general abdul qassim kassem, hussein kamel, terrorist, missil, saddam husseins, terro, intern

次に主張語の差分だけを用いた場合の結果をこれも人手で要約する:

クラスタ 1 : 主張語の差分

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, terrosist, 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る。ラベルとして"terrorist"、"nuclear weapon"、"chemical biological weapon" などの単語が現れていることから、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。次に、主張語の差分の結果と比較すると、主張語の差分では個々の単語として得られた"nuclear" "weapon" "inspector"が提案手法で単語の並びとして抽出している。このことは、高度な意味の把握を可能にしていると言える。

同様に、クラスタ 2 (Start War) は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る。ラベルとして"osama bin laden"" mass""destruct""claim"、報道関係の言葉が現れている。

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

クラスタ 2 : 提案手法の結果

武装: army, time, attack, coalit, attempt
国際: world
アメリカ国内: leader, american, claim
イラク国内: osama bin laden, iraqi president saddam hussein, dictator, author, party
報道: report, live, fact
UnitedNations: mass, destruct

クラスタ 2 : 主張語の差分

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, BinLaden, Amicu, Dictator party
報道: report, talk, live, fact
UnitedNations: WeaponMassDestruction, Answer

このクラスタではあまり両手法で得られたラベルに差異は見られない。主張語の差分でも、”bin laden” ”weapon mass destruction”の言葉が現れる。

クラスタ 3(`Uday and Qusay`) は以下のようなラベルを得た。

クラスタ 3 : 提案手法の結果

武装: military, oper, fire, troop, milit
ウダイとクサイ: husseins son udai qusai, july, udai hussein, qusai hussein, family
イラク: baath party, intellig, secure, power, intelligence, order, iraq war, foreign newspap dare attack
アメリカ: capture suddam hussein start, american troop,coalit force, secure force, america, fought saddam dictatorhip, expens endeavor time rebuild iraq

クラスタ 3 : 主張語の差分

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 3 はウダイとクサイの死亡した時期であり二人に関連した言葉が現れる。更に、単語の並びを考慮する手法では、ウダイとクサイに関連す単語の並び”husseins son udai qusai” や、フセインの捕獲作戦を意味する”capture suddam hussein start”単語の並びが現れている。これらのラベルは、(すくなくとも単語よりは) 高度に抽象化されており、内容をより的確に表現するものとなっている。

クラスタ 4 (`Saddam Captured`) は、フセインが捕まった時期である。

クラスタ 4 : 提案手法の結果

武装: soldier
国際: arab, world
アメリカ国内: president bush, nation
UnitedNations: weapon mass destruct
フセイン: captur saddam hussein, war crime, president saddam hussein, intern, trial, tikrit
イラク体制: iraqi govern council, war iraq, regime
報道: inform, public

クラスタ 4 : 主張語の差分

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: capture, family, sunday, death, trial, hole, crime, tikrit
イラク体制: administr, govern, leader, nation, coalit, regim

クラスタ 4 での大きな出来事はフセインの捕獲である。主張語の差分だけを用いた手法でも”capture” ”tikrit”の言葉や多くの報道関係の言葉が現れているが、特に提案手法で現れる”capture saddam hussein” は、主張する意図を高度に抽象化したものとなっている。

クラスタ 5(`after getting Saddam`) ではその後の状況変化を捉えた語が現れる。

クラスタ 5 : 提案手法の結果

United Nations: red cross visit saddam hussein, mass destrct, unit nation, author
報道: forc kill prove loyalti, rememb kill stop al quaeda, hussein act hitler gass people, escap prison continu work praty,
アメリカ: state, america, framework sanction committe full approv, lead,
国際: middl east
武装: military

クラスタ 5 : 主張語の差分

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

主張語の差分でバラバラに現れた”visit” ”red cross”は提案手法では”red cross visit saddam hussein”と現れる。このラベルも、これまでと同様に、意図を高度に抽象化したものとなっており把握しやすいと考えられる。

5.5 評　　価

これらから判断し、得られたラベルは予め与えた解釈と対応している。STC を行う為の手順から得る単語の並び、および KeyGraph から得る主張語の双方を考慮した手法は、内容を高度にかつ的確に表すラベル付けが行えたことを表している。利

用者は、時制クラスタの内容を把握するために、自動的に抽出されたこれらのラベルを探索することにより、クラスタが表現する事象を即座に理解できるであろう。

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 STC を行う為の手順に基づく本手法が、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。

6. 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、Suffix Tree Clustering を行う為の手順に基づく手法で単語の並びを抽出し、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手法を提案した。

最初に、各クラスタで単語の並びを Suffix Tree Clustering を行う為の手順に基づいて抽出、各クラスタの重要語を KeyGraph に基づいて抽出した。次に、主張語を考慮し単語の並びより各クラスタにラベル付を行った。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、クラスタに対して高度な意味の把握が可能であることを意味している。ラベルとして得られた単語の並びを抽象化・集約化ができるならば、その可能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。

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] 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] 大沢幸生 : KeyGraph　ー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

[11] Oren Zamir and Oren Etzioni.: Web Document Clustering: A Feasibility Demonstration, SIGIR 1998: 46-54

[12] 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