時事性を考慮したWebからのトピック検出と追跡と要約に関する研究

森, 正輝 / MORI, Masaki

(発行年 / Year)
2006-03-24
(学位授与年月日 / Date of Granted)
2006-03-24
(学位名 / Degree Name)
修士(工学)
(学位授与機関 / Degree Grantor)
法政大学 (Hosei University)

2005年度
修 士 論 文

時事性を考慮したWebからの トピック検出と追跡と要約に関する研究

STUDIES ON TOPIC DETECTION AND TRACKING AND SUMMARIZATION FROM WEB BY USING TOPICALITY

指導教官　三浦 孝夫 教授

法政大学大学院工学研究科
電気工学専攻修士課程

04R3246　森 正輝
Masaki MORI

目 次

第1章　序論

1.1　問題の背景

現代のようなインターネット社会では，Web やディジタルアーカイブデータを介して，文章，画像，映像，音楽などの多種多量のデータが簡単に手に入る．どのような内容であろうとも，辞書や事典の変わりにネット検索することが当たり前になり，誰でも容易に興味のある情報を得ることができるようになった．更に，Web は驚異的な速度で成長している．毎日何百万ものページ，毎月数百ギガのページが更新され，ページ数も驚異的に増加し続けていると言われている．しかし，多くの図書と違って，相互の関連性に乏しく，内容の信頼性が保障されず，未整理のままに放置されるという状況にある．言い換えれば，形式化せずにかって気ままな意味づけを与え，内容の整合性になんら制約されないままの状況であると言える．近年のこの状況は 60 年代から 70 年代と極似している．未整理のまま統合されず相互に矛盾を有する重複データが点在していた時代に，データベース技術が確立された．技術の根幹には重要なアイデアが含まれる．データが有する意図あるいは特性を予めスキーマ情報として管理システムに蓄え，これを用いて正確かつ効率よくデータを処理するという考えである．

現在直面する問題の多くはかつて扱ったものと類似しているが，対象となるデータに含まれるスキーマ情報は，未知あるいは未整理でありかつ部分的・主観的である．このような非定型性が，従来のデータベース技術の適用を大変困難にしている．対象となるデータを管理するためには，内容が意図する多様性を捕らえ分類格納する技術あるいは分散管理するモデルが必要である．ここでは，的確かつ拘束に検索する検索能力や，部分的にしか定まっていないデータを強調的に意図を確率していく様相性も要求される．しかし，最も難しい問題は，データの意図をスキーマとして与える方法や抽出する方法がないことである．

典型的なデータをニュース文書に見ることができる．新聞や TV などで流されるニュースは，発信される時間を伴う記事の流れ (時系列ストリーム) であり終わりのない文書である．各記事は主語述語など比較的形式の整った文章から構成されるが，何に関する話題か (トピック)，どんな事件を扱っており，どこが強調されるべきかなど，注釈や要約が付いている事は少ない．記事間の関連，事件やトピックの切れ目も明確ではない．

例えば，次のニュース (1990 年代後半) を考えた場合：

1. 英国トニーブレア首相が香港で生卵をぶつけられた一連の事件報道

2. 香港の反英感情の高まりと，中国復帰に向けた政治的動きの報道

3. ブレア首相はその後日本を訪問し，首相会談を終えて帰国との報道

各記事はどれかの内容と密接に関わっている．実際，時間順に報道記事を並べてみると，意外に単一の事件にかかわるものが多く，人間にとって整理しやすいことが経験的に知られている．これら事件を追跡し相互の関連性 (流れの合併・分割・相互影響・復活など) を捕らえれば，全体を理解する上で重要な働きとなる．実際，上述 1,2,3 は「英国の東アジア重視」というトピックで語ることができる．

これらのデータを計算機で処理しようとすると，人間にとって簡単な内容把握が，困難になる．例えば，生卵事件は「東アジア重視」には影響しない．見出しを「背広の胸元が汚れた」とするニュースが，事件あるいはトピックとどう関連するかを判定する要素にはならない．どう表現すべきなのか，その手がかりもない．

1998 年ごろから米国で開始された Topic Detection and Tracking（トピック検出・追跡）プロジェクトは，ニュースストリームの内容を解析し，どのようなトピックに関するのか，あるいは関連しあう記事は何かといった情報を，データ自体から抽出する技術の確立を目指している．複数言語を多様なメディア (配信記事，放送あるいはそのトランスクリプション) で表現し，トピックや事件を抽出しようとする研究がその中心にある．ひとつのトピック内では，各文書を単語ベクトルで表し時間軸上に配置してクラスタリングした結果は，個々にひとつの事件と対応することが知られている．これにより，人間の経験が計算機処理で立証できる例である．でも，どのようなトピックがどの事件とどう関係するのか，更にはその内容の要約・ラベル付けまでを実行しないと，人間には理解できない．

ニュースストリームに対する技術は，Web 検索に直接応用できる．私たちは，普段，検索エンジンを使い検索語を与えて，3 億ないし 30 億の URL データベースから，いくつかのトピックに関する Web ページを検索する．しかし，利用者は膨大な検索結果を全て調べることはせず，最初の 10 又は 20 ページだけで興味のあるページを探し出すと言われている．いくら詳細に探索しても，検索結果を一見しただけで内容を把握できず，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困難である．もし Web ページ内容が時間的に捕らえられれば，TDT 技術をそのまま利用できる．検索結果を意味的に分類して事件を抽出し，これらを追跡することでトピックを見出し，内容を要約できたならば，検索結果をより効果的に吟味することができ，負担も軽減できることは明らかである．

1.2 定義

本研究で使用する**事象**を定義する．本研究では事象を「 ある一定の時間，特定の場所で発生した事件，事故，出来事」と定義する．事象の例とて，9/11 の同時多発テロ，アフガニスタン全面攻撃，首脳会議，タリバーン統治の征服などの事件，出来事が挙げられる．

次に，**追跡**を定義する．9/11 の同時多発テロ，アフガニスタン全面攻撃，首脳会議，タリバーン統治の征服などの事象は発生した時間，場所も全く違うが 9/11 同時多発テロから連なった一連の流れを事象の関連性から見出すことができる．検出された事象間の関連性を見出し，関連性があれば事象を辿ることを本研究では**追跡**と定義する．

最後に，**トピック**を定義する．9/11 の同時多発テロからの一連の事件，出来事である，アフガニスタン全面攻撃，首脳会議，タリバーン統治の征服，世界中での小規模テロの発生などの事象は追跡することが可能である．つまり，これらの事象は個々に独立した事象ではなく，1 つの大きな話題に関する事象の集合と言える．この大きな話題を本研究では**トピック**と定義する．

1.3 扱う問題

本研究では，利用者が検索エンジンの検索結果から利用者がトピック全体を容易に把握できることを目指し，検索エンジンに検索語を与えて得られた Web ページ集合に対し，事象の検出，要約，追跡を行う．検索エンジンを用いて得られた Web ページは 1 つのトピックについて論じられた Web ページ集合であり，事象がどのトピックであるか特定する必要はない．

最初に検索エンジンの検索結果から事象の検出を行う．事象の検出を行うことで，どのような事件，出来事が発生したかを判断できる．次に，各事象に対してラベル付けにより要約を行う．要約を行うことで，人間が事象を理解することが非常に簡単となる．最後に，追跡を行う．これにより，事象同士の関連性を明らかにしトピック全体を把握しやすくなる．

1.3.1 事象の検出

TDT の分野において，時間軸で文書をクラスタリングすることで各クラスタが事象に対応することが知られている．しかし，Web ページにはページの内容を表した時間は明記されていなく，Web ページから得られるいくつかの時間情報から内容の示す時間（有効時間）を推定する必要がある．有効時間の決定方法を決定することができれば，時間軸で Web ページをクラスタリングすることで事象を得ることができる．

本研究では，URL の一部に暗示された時間，受信ヘッダに明記された時間，Web ページの文書の直前に現れる時間より有効時間の推定を行う．有効時間の推定を正確に行うことができれば，得られたクラスタは何かの事象に対応している．各クラスタを検証し事象と対応していることを実験により証明し有効時間の決定方法を決定する．

1.3.2 要約

要約には，文章の中から要約に適した文を抜粋する圧縮と，単語，単語列を抽出するラベル付けがある．

圧縮では圧縮のレベルは任意に選べる．100 の文から構成される文書に圧縮率 10 %で圧縮を行った場合，文書から 10 の文が抜粋される．圧縮率が高いほど，抜粋される文の数は減少するのだが，”he”，”they”などの代名詞が参照している名詞がわからなくなったり，前後の文の繋がりを無視した文の抜粋を行ってしまう．他にも，箇条書きで「反逆者は 3 つの要求を行った」とあり，3 つの列挙のうち，2 つだけを要約として抜粋した場合，読者に誤解を与えてしまうことになる．圧縮率が低い場合，多くの文を抜粋することになり利用者の負担を軽減させることにはならない．

ラベル付けでは，頻度の高い単語をラベルとする方法が考えられる．しかし，頻度の高い語をラベルとした場合，頻度の非常に高い”the”，”I”，”a” などの不要語を取り除いたとしても，ラベルとしては適さない語が抽出されてしまうことが知られている．特に，類似度の高いトピックではラベルに大きな差異はなく，検索エンジンから得られた Web ページ集合に対して語の頻度に依存したラベル付けは適当な手法とは言えない．更に，利用者にクラスタの内容を把握しやすくするには，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良いことが知られている [11]．

本研究では，KeyGraph の考え方を用いて，語の共起関係より Web ページの著者の主張点を表した語を，Suffix Tree を用いて語の並びを抽出し，クラスタの時制的な側面を考慮したラベル付けを行う．

1.3.3 追跡

検索エンジンによって得られた Web ページ集合は複数の事象が混在している．クラスタリングによって得られた事象同士の関連性が理解できなければ，利用者はトピック全体を把握することはできない．すなわち，正確な追跡が行えれば，利用者がトピック全体を把握することが容易になる．

追跡の手法として，分類法を用いた追跡がある．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を判断し追跡を行うものである．しかし，時間の経過にしたがって，” 関連性の薄かった事象が合併 ”” 1 つの事象が分離 ”” 新たなトピックの発生 ”” トピックが消滅 ” するため，初期値に依存する分類法では，トピックの概念の変化に対応できない．

本研究では，トピックの概念の変化に対応するために，KeyGraph の考えを用いて各事象の概念を抽出し，古い概念を捨て新しい概念を取り入れ事象の追跡を行う．

1.4 関連研究

1.4.1 Web 検索

Yahoo!，Google などの各検索エンジンは，クローラーと呼ばれる自動巡回ロボットを使用しリンクを辿ることで Web ページの収集を行い，3 億ないし 30 億の 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に検索語を与えて検索を行うことで利用者は，検索語を Web ページに含む，Web ページの URL を見つけることができる．検索結果はランキング形式で表示される．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て結果を表示するいくつかの手法が提案されている．非常に長い検索結果を絞り込むために，検索エンジンに対して利用者は論理質問を使用することができる．実際には，利用者本人が興味あるトピックを言葉で表現することは非常に難しく，論理質問を使いこなして検索を行える利用者は少ない．

他の Web 検索手法として，ディレクトリィ検索と呼ばれる検索手方法がある．予め用意されたカテゴリに人手によって Web ページを分類しておき，利用者がカテゴリを辿ることで，興味のある Web ページを見つけるという検索方法もある．しかし，驚異的に増加している Web ページを全て人手で分類することは不可能であり，分類の基準も主観に依存したものになってしまう．更に，古いカテゴリが消滅したり，新しいカテゴリが発生したり，カテゴリが分離，合併することもあり，カテゴリそのものの日々の変化に対応する必要もある．

1.4.2 自動要約

近年，我々は非常に高速な移動体通信と膨大な情報の蓄積へのアクセスを手にした．しかし，膨大な情報を 1 人の人間が理解するには非常に時間がかかり，効率的な作業であるとは言えない．それ故，自動要約は必要不可欠な技術である．自動要約の分野は 1950 年代から存在する分野で，自然言語処理，情報検索，図書館学，統計，認知心理学，人工知能などの分野から自動要約の手法が提案され議論されている．この分野は要約手法の評価の問題を必ず引き起こす．1 つの原文から，複数の要約が可能であり，どれが望ましいと言う明確な基準がないことが問題である．要約器によって得られた要約は，処理した原文の分野の専門家又は，著者でない限り，直感的に良い要約を選ぶことは困難である．評価についての議論も，この分野では非常に重要である．

1.5 論文の構成

本研究では，以上の問題について以下の構成で論じる．第 2 章では，Web からの時制クラスタの抽出，KeyGraph を用いた事象の検出までを論じる．第 3 章では，時制クラスタに対して，Suffix Tree Clustering を用いたラベル付け論じる．各時制クラスタ

に対して抽象度の高い要約をラベル付けにより行う．第 4 章では，時制クラスタから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ながら追跡を行う．第 6 章で結論とする．

1.6 発表論文

1. 森正輝, 三浦孝夫: “Web ページの時間順序付け”, 2004 年総合大会春季, 電子情報通信学会, 2004.

   Web ページの有効時間の推定方法の有効性を実験により示す

2. 森正輝，三浦 孝夫，塩谷 勇: “Web からの時制クラスタの解釈”, 日本データベース学会 Letters (*DBSJ Letters*) Vol.3, No.2, pp. 109-112, 2004

   Web ページの有効時間の推定方法，事象の検出手法を提案．実験により有効性を示す．

3. Masaki, M. Miura, T. and Shioya, I.: “Extracting Event from Web Pages”, *International Conference on Advances in Intelligent Systems: Theory and Applications (AISTA 04)*, (CD-ROM), 2004

   Web ページの有効時間の推定方法，事象の検出手法を提案．実験により有効性を示す．

4. 森正輝, 三浦孝夫, 塩谷勇: “Suffix Tree Clustering を用いた Web ページ集合のラベル付け”, データ工学ワークショップ（*DEWS*）, 2005.

   事象に対して主張を表した単語，単語の並びを考慮することでラベル付けにより要約を行う．実験により有効性を示す．

5. Masaki, M. Miura, T. and Shioya, I.: “Abstracting Temporal Clusters”, *Internet Technologies and Applications (ITA 05)*, (CD-ROM), 2005.

   事象に対して主張を表した単語，単語の並びを考慮することでラベル付けにより要約を行う．実験により有効性を示す．

6. 森正輝, 三浦孝夫, 塩谷勇: “時制クラスタのトピック追跡”, データ工学ワークショップ（*DEWS*）, 2006.

   初期値に依存しない，追跡手法の提案をする．事象の概念を抽出し，事象同士の関連性により追跡を行う．実験により有効性を示す．

第2章　Webからの時制クラスタの解釈

2.1　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ある．このため Web ページ集合の内容を素早く容易に把握するための研究が近年注目を浴びている [2, 12, 13].

本稿では Web ページ集合からの自動的な事象抽出手法を提案する．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のを手助けする為に，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，新たに検索結果のリストが長くなっ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 ページの中から有益な 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出力されるべきである．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．

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, 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．

ここでは，全ての情報が明示的もしくは暗黙のうちに時間の情報を持つと考えられ

る．時間の情報なくしては成り立たず，ここで扱うWebページに事象を検出するためにタイムスタンプを推定することは必要不可欠である．Webページに信頼性のあるタイムスタンプを類推できれば，事象又は傾向をより簡単に検出することができる．

本稿では，URLの検索結果の並びを無視し，各Webページのタイムスタンプの類推を行う．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) に従って理解されるが，必ずしも文書の内容時間が文書の作成・修正された時間，すなわち**作成時間** (creation time) や**トランザクション時間** (transaction time) と一致するものではない．

有効時間を類推しWebページを時間軸にしたが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．本稿の，基本的な考えは，TDTのようにWebページをクラスタ化する観点から，適当な方法で事象を抽出することである．さらに，本稿では，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る．

本稿では，２章でどのようにWebページから有効時間の抽出を論じる．３章で事象の抽出方法，その有効性を論じ，Googleを使い実験的な結果を論じる．4章で得られた事象を解釈するキーワードの決定の方法を論じ，5章で実験結果の考察を行い，6章で結論とする．

2.2 有効時間の推定

本研究では文章の文法を解析することなくWebページから有効時間を類推する方法を提案する．ここでは3種類の時間を考慮する．(a) コンテンツに明示的に現れている内容時間 (CT)，(b) 作成時間 (UT)，(c) 更新時間 (TT) である．(b) はURLの一部に含まれており，(c) は受信ヘッダに明記してある．有効時間とは，Webページが示そうとしている時間を意味する.「松井稼頭央が2003/12/09にNew York Metsに入団」という文を例にとると，内容時間はWebページの文章に明示的に出現しているタイムスタンプである．この場合「2003/12/09」である．作成時間はWebページが生成された時間を言う．ここでは「2003/12/10」と仮定する．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．更新時間は，Webページが格納された時間又は，最後に更新された時間を言う．これを「2003/12/11」と仮定する．各Webページを解析し，内容時間，作成時間，更新時間を全て抽出しどれが有効時間により近いかを調査する．内容時間はそれぞれの文章の最初の文の前に現れるもので,「Jan 04, 2004」又は「January 3, 2004」のようなパターンのものを抽出する．複数の内容時間が抽出できる場合は，すべてを抽出する．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．

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

tech/main591195.shtm

最初の URL は 2003/01/05 の作成時間を含んでおり，次も同様に 2004/01/04 の作成時間を含んでいる．しばしば，URL は作成時間 (UT) を含む．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 のヘッダーが含まれる場合，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．だが，すべての Web ページの CT，UT，TT 又は VT を必ず含むわけではない．有効時間を類推する為に，抽出した CT，UT，TT のどれが VT に近いか調査する．その為に，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が有効時間を調べる．この時，全ての Web ページが常に CT，UT，TT，VT を含むわけではないことに注意する．例として，「under construction」が有効時間を持たなくても，UT 又は TT を持つ場合がある．このとき，null を用いて表す．一方，複数の CT を抽出できる Web ページでは各 CT のページのタプルを生成する，すなわちどのページも 1 つ以上の CT を持つことになる．

まず最初に，テストページ p の集合 T を取得する：

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

$T_P = \{p| < p, .. >\in T\}$ と定義する．

$p \in T_p$ が与えられたとき，ET(p) を推定するために，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：

$$\begin{aligned}
V &= \{< p, VT(p) > | p \in T_P, VT(p) \neq null\} \\
P_C &= \{< p, CT(p) > | p \in T_P, CT(p) \neq null\} \\
&... \\
P_{CT} &= P_C \cup \{< p, TT(p) > | p \in T_P, CT(p) \\
&\quad = null, TT(p) \neq null\} \\
&... \\
P_{CTU} &= P_{CT} \cup \{< p, UT(p) > | p \in T_P, CT(p) \\
&\quad = null, TT(p) = null, UT(p) \neq null\} \\
&...
\end{aligned}$$

ここで V は全ての可能な答を意味する．他の定義はどのよう推定時間 (ET) を得るかを示す．例えば，P_{CU} は CT が null でない限り内容時間とし，$CT(p)$ が null だが $UT(p)$ が null でないときは $UT(p)$ を内容時間として類推する．この意味で，P_{CU} は内容時間の類推方法を示し，これを CU と示す．

Ans(答)，Rec(再現率) と Pre(適合率) を次のように定義する：

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

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

$$Pre = |Ans(P)|/|P|$$

Rec は，どれだけの答を $Ans(p)$ がカバーできたか，Pre はどれだけ正解を $Ans(p)$ が含んでいたかを示す．本稿では以下の式で示される F 値を用いる:

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

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 ページから類推時間を得る方法を求めたことになる．

本実験では，Google で `Kazuo matsui` を検索し，Top300 ページを得た．そこからリンク切れ，Weblog 以外の 235 の URL を対象とし，これらのページに対して手動でタイムスタンプを決定し 211 ページのタイムスタンプを得た．そして，各 Web ページから内容時間 CT，作成時間 BT，更新時間 TT の抽出を行った．スキーマは，内容時間の類推方法を示し，ExpTime は null ではない時間を持つページの数であり Ans はスキーマごとの答の数を示す.

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

この結果からわかるように，UC が Pre 値が最も高い値であるが，F 値が最大となるスキーマは CU である．CU は，多くの答をカバーしている事を意味する Recall の値が最も良い．一方 CUT の F 値は TT の Pre 値が非常に悪いために小さくなっている．

2.3 事象の抽出

Web ページ集合から有効時間 (VT) を推定したが，次に事象抽出を行う．この章では，類推した VT を用いて Web ページの事象 (Event) を抽出する方法を提案する．

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．以下では Web ページ集合を検索エンジンに検索語を与え，その結果を得ており，Web ページ集合は 1 つのトピックについて論じられていると考えられる．

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．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．以下はその結果である．

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

図 2.1 に結果を示す．結果として，8 つのクラスタを得た．半分のクラスタは，クラスタの構成要素が 5 以下であり非常に小さいので無視する．残る各 4 つのクラスタを解釈する為に，`Kazuo Matsui` を含む文を選択して，頻繁な語，特徴的な語を手作業でを取り出した．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 major   leagues, 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
http://www.taipeitimes.com/News/sport/archives/2003/
12/12/2003079339/print
```

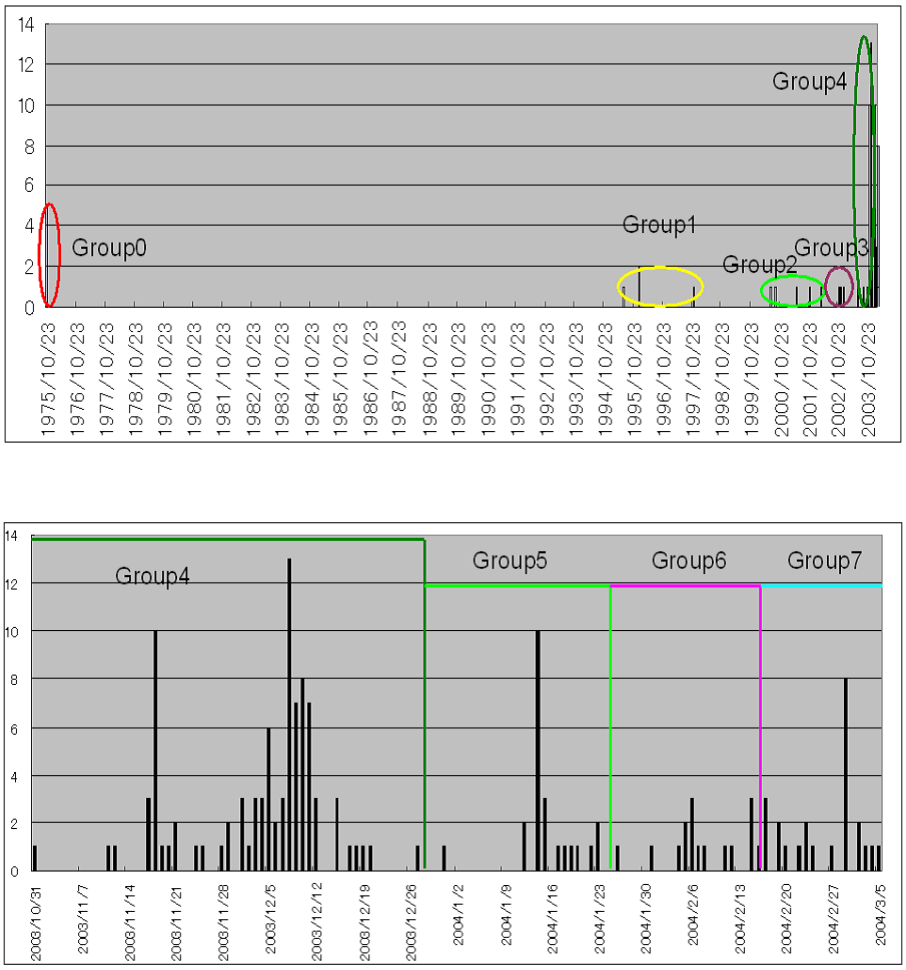

図 2.1: Clustering Kazuo Matsui pages

```
  NY Mayor welcomes Matsui No. 2
....
```

これらのクラスタの文章より，このクラスタを Mets welcome sign major league と我々が解釈した．全てのクラスタの解釈を下に示す．

```
(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
```

すべてのクラスタの解釈は Kazuo Matsui に関したいくつかの事象で適当なクラスタであり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る．自生的な側面を持つ Web ページにも TDT と同じ傾向を持つことを示している．

2.4 事象の解釈

Web からの時制クラスタを自動解釈する試みを提案する．本稿では，各クラスタから重要な語句を抽出し，これをラベルとする方法をとるが，ここでは KeyGraph の考え方を用いる．

KeyGraph とは，文章中に出現する単語の出現頻度と共起関係から文章の主張点を把握し，キーワードを抽出する手法である [10]．KeyGraph では，文章中に頻繁に出現する言葉は文章が書かれる上での前提条件，つまり基本的な概念であり「土台」と呼び，更に「土台」によって支えられている語が文章の「主張」(筆者の主張) であると見なす．KeyGraph の生成は，いくつかのステップからなる．まず，文章から不要語を取り除き，上位定数個の頻出単語間の共起度を計算し語同士の間に定数個の共起リンクを張る．次に，共起度の薄いリンクを切断して土台となる語のグループを作る．土台の語グループと，不要語を取り除いた文章中に出現するすべての語の共起度を算出し，値の高い定数個の語 (主張) の共起リンクを残す．更に，共起リンクの設定された，土台となる語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．これを土台と主張を結びつける重要語とみなし上位語を選定する．

本稿では，この KeyGraph により得られる主張語を時制クラスタの解釈に用いる．実際，時間軸でクラスタリングされた Web ページは相互に類似性が高く，得られた主張語集合には極端な差異は生じない．ただ，時間軸に沿って変化している様をとらえるため，1 つ前のクラスタの主張語集合と比較し，その差分で時制クラスタを解釈する．このとき，時間軸で一番古いクラスタは差分計算ができないので，以下では，主張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．

2.5 実験

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．先に述べた様に，最初に URL のリストを取得し，提案した手法 CU にしたがって VT の類推を行う．

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，669 のページを得た．この 669 ページを時間軸によりクラスタ化すると，図 **??**で示すような 6 つのクラスタを得た．

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

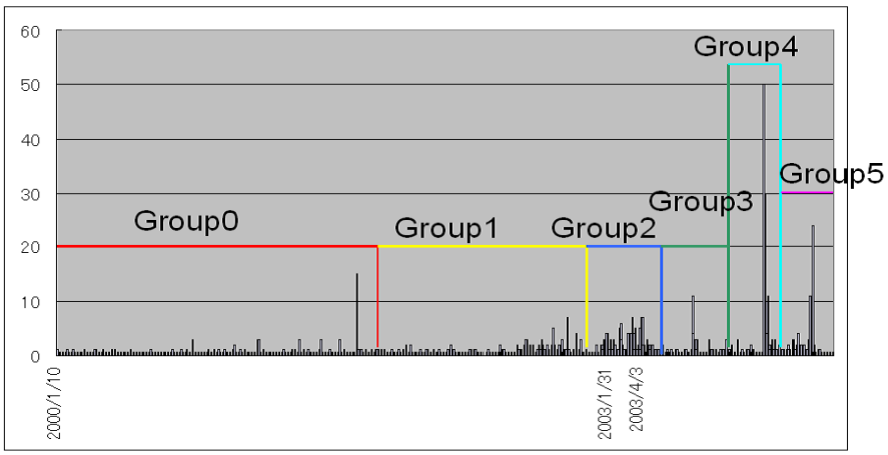

図 2.2: Hussein のクラスタリング結果

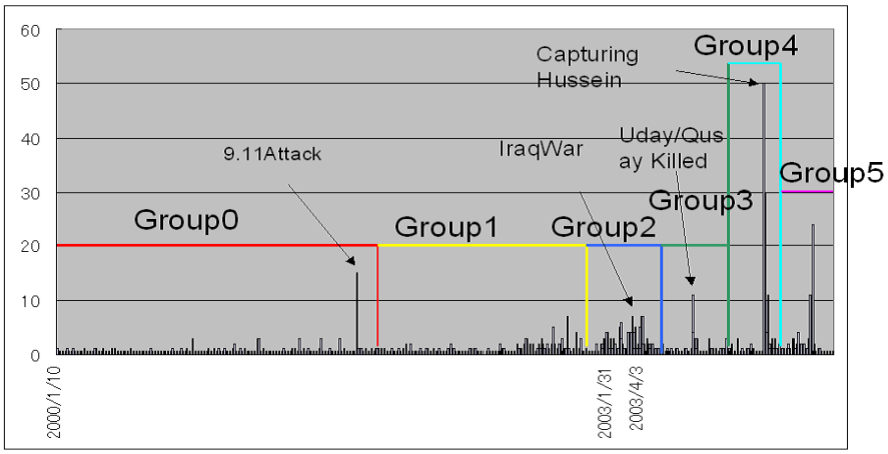

図 2.3: 実際の事件と時制クラスタの対応

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
IRAQ: CRIMES AGAINST HUMANITY
Leaders as Executioners
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．

次に，上述クラスタを KeyGraph 手法を用いて主張語を取り出し，差分を抽出する．

クラスタ	主張語
0	31
1	50
2	54
3	40
4	63
5	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, christianscienc-moni, tore, claim, control, defect, march, missil, opposit, terrorist, plan, terror, senat, agreement,

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，さらに得られた主張語集合は，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[1]．ここでは，これを次のように人手で要約する:

[1]たとえば Wordnet などの辞書を活用すればよい．

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

これらの内容は，上述の人間による解釈 (`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`) は次のような主張語と対応している．

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, BinLaden, Amicu, Dictator party

報道: report, talk, live, fact

UnitedNations: WeaponMassDestruction, Answer

クラスタ 3(`Uday and Qusay`) も同様に次のような主張語と対応する．

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`) では特に報道分野の語が現れる．

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`) ではその後の状況変化を捉えた語が現れる．

往来: visit, com

支援・体制: redcross, author , ICRC

UnitedNations: ICRC, evid

これらから判断し，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．以上から，提案手法が我々の知識で有効性が示せたと言える．

2.6 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自動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類推を行い，小さなテストページ集合で予備実験を行い，経験的に類推方法 P_{CU} を採用した．次に，K-meansアルゴリズムによりクラスタを作り，KeyGraphの方法に基づいてそれらの解釈を行った．実験に基づく結果は，時制Webページで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．有効時間を類推できれば，事象の検出と追跡も容易なると考えることができる．

第3章　Suffix Tree Clusteringを用いたWebページ集合のラベル付け

3.1　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, 12, 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, 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 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 Web ページ集合に Suffix Tree Clustering (STC) に基づく手法を用い，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．

3.2 ラベル付けの意義と目的

3.2.1 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．Web ページ集合の内容が，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 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 [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 を用いることで抽出しラベル付けを行う．

3.2.2 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[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[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[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．

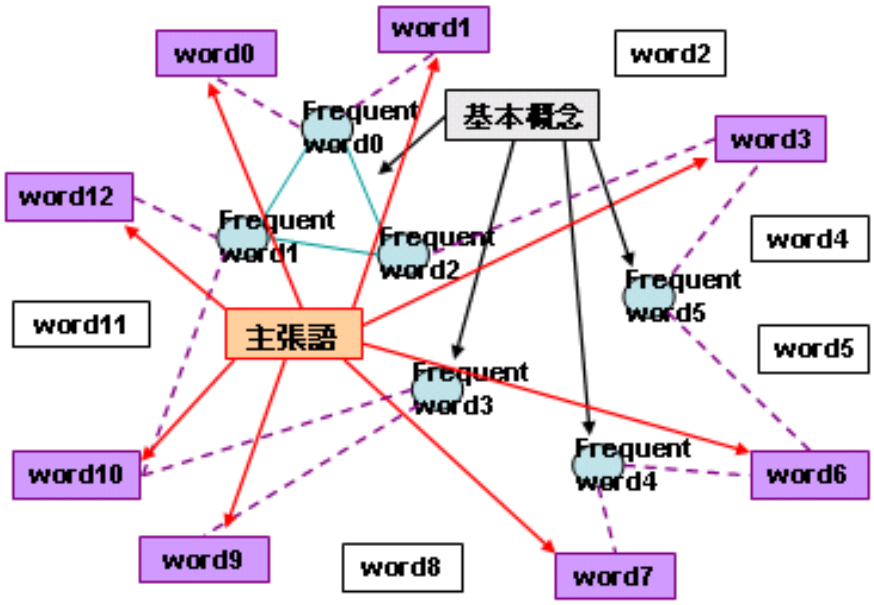

図 3.1: 基本概念と主張語

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 3.2 に例題の KeyGraph を示す．

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．

[1]KeyGraph では「土台」と呼ぶ．
[2]KeyGraph では「屋根」と呼ぶ．
[3]KeyGraph では「柱」と呼ぶ．

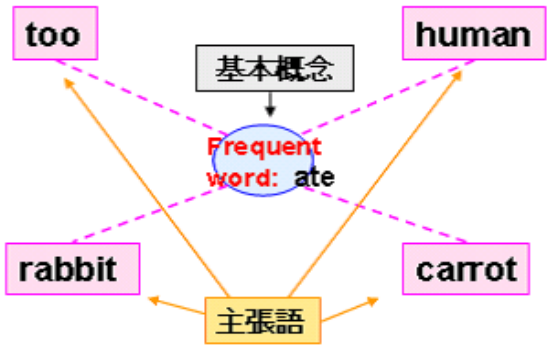

図 3.2: 基本概念と主張語

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．

例題 2 以下に示す 3 つの文書に対して STC を行う．

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．

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 3.1: 各ノードの詳細

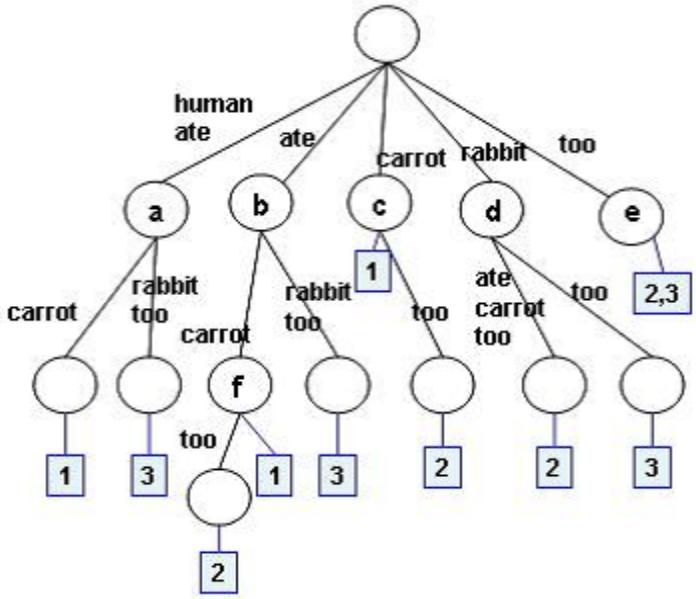

図 3.3: Suffix Tree

STC では，各ノードごとにノードを構成する単語数とノードが発生する文書数を用いて各ノードのスコアを算出する．一定値以下のノードを切り捨てた後，2 つノード間で，ノードが発生する文書が半数以上共通する場合，ノード間にリンクを設定し，リンクで繋がる，ノードをクラスタとみなしクラスタ化する．この手法は，Single-Link Clustering と同等なものでありクラスタリング手法として適当である [11].

例題では，ノード a,b,c,d,e,f からなる 1 つのクラスタを得た．仮に，ノード b が不要語であったならば，3 つのクラスタを得ることになる．

3.3 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

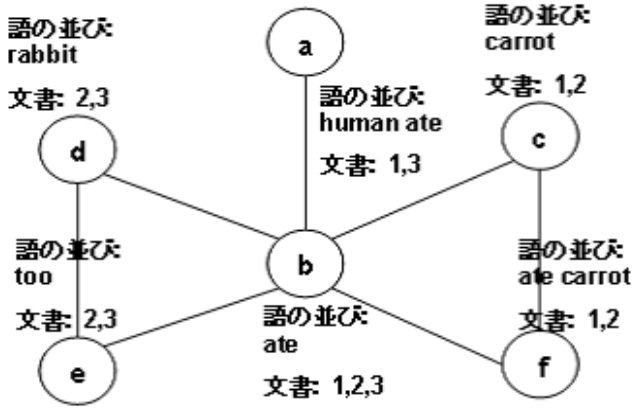

図 3.4: クラスタリング結果

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．

3.4 ラベルの決定

3.4.1 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．

3.4.2　単語の並びの抽出

STC 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．

3.4.3　ラベルの決定

KeyGraph に基づく主張語，STC 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 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 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．

3.5 実験

3.5.1 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．

3.5.2 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

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
```

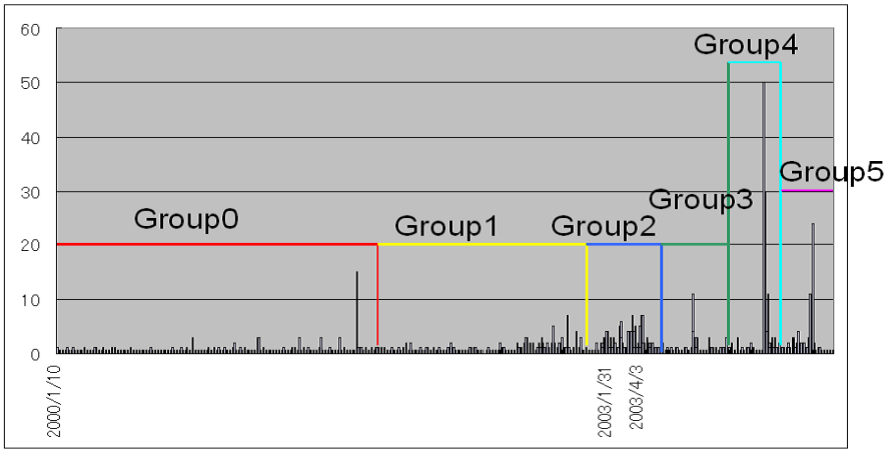

図 3.5: Hussein のクラスタリング結果

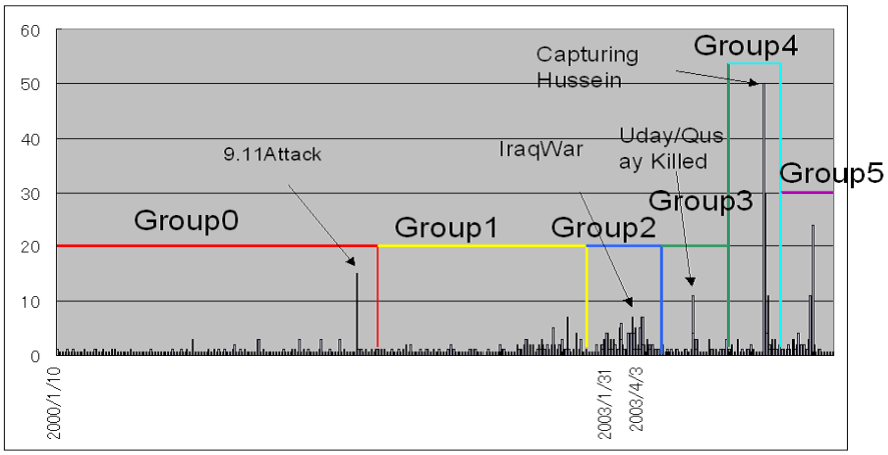

図 3.6: 実際の事件と時制クラスタの対応

```
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 4.7 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．

3.5.3 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 3.2: 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).

3.5.4 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

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 3.3: ラベル

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[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

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

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．

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る単語の並び”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．

3.5.5 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 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 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．

3.6 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 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 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．

第4章　時制クラスタのトピック追跡

4.1　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, 12, 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．

4.2 トピック追跡とラベル付けの意義と目的

4.2.1 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 [11].

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．

4.2.2 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[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[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[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 3 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 4.2 に例題の KeyGraph を示す.

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．

[1]KeyGraph では「土台」と呼ぶ.
[2]KeyGraph では「屋根」と呼ぶ.
[3]KeyGraph では「柱」と呼ぶ.

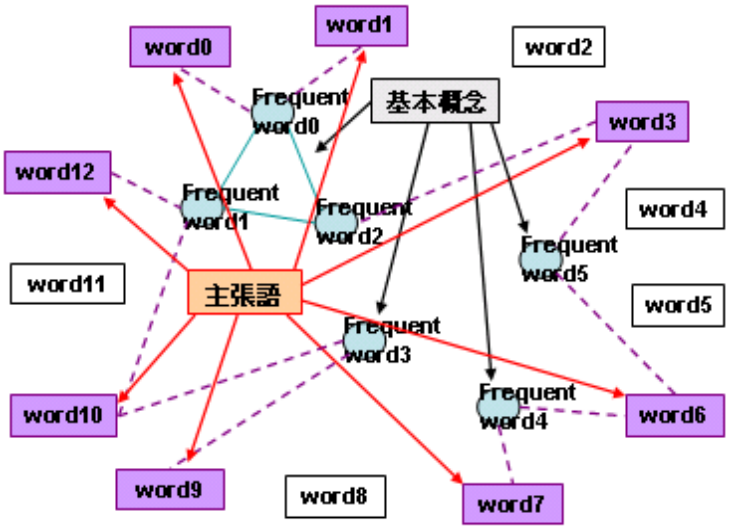

図 4.1: 基本概念と主張語

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 4 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 4.1).

4.3 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].

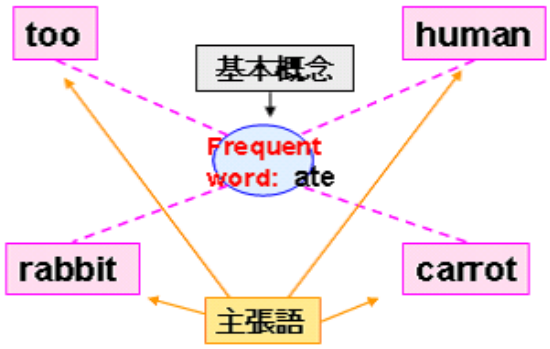

図 4.2: 基本概念と主張語

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 4.1: 各ノードの詳細

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．

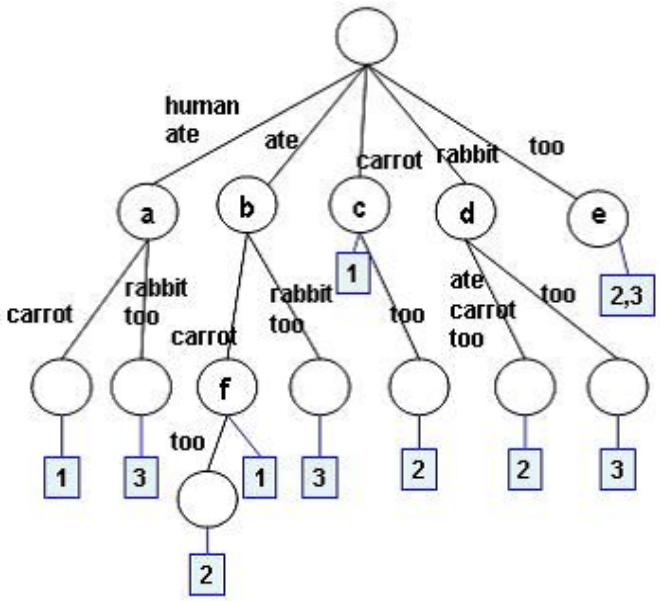

図 4.3: Suffix Tree

4.4 トピック追跡

4.4.1 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．

4.4.2 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 4.4）．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．

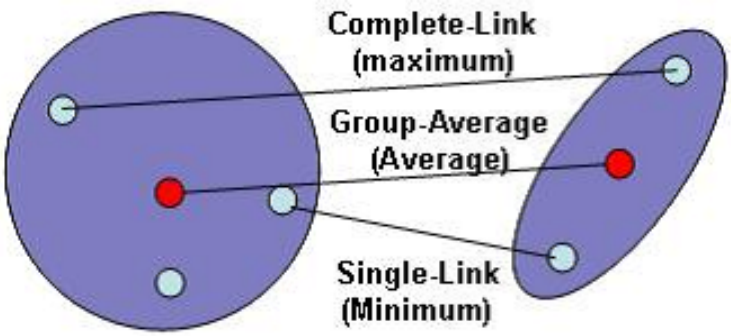

図 4.4: Complete-Link クラスタリング

クラスタリングを行った後，基本概念を抽出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．

4.4.3 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，古い概念を捨てた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が可能となる（図 4.5）．

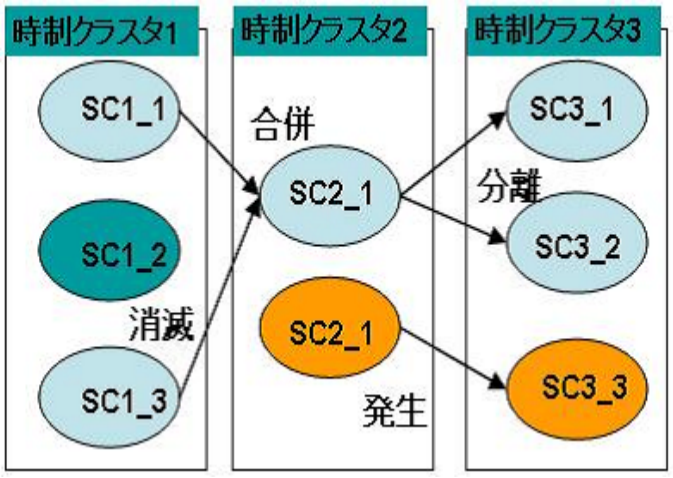

図 4.5: Topic の消滅，発生，合併，分離

本稿ではサブクラスタの要素の平均をサブクラスタの代表点とし，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．

4.5 ラベルの決定

4.5.1 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．

4.5.2 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．

4.5.3 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].

4.6 実験

4.6.1 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．

4.6.2 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

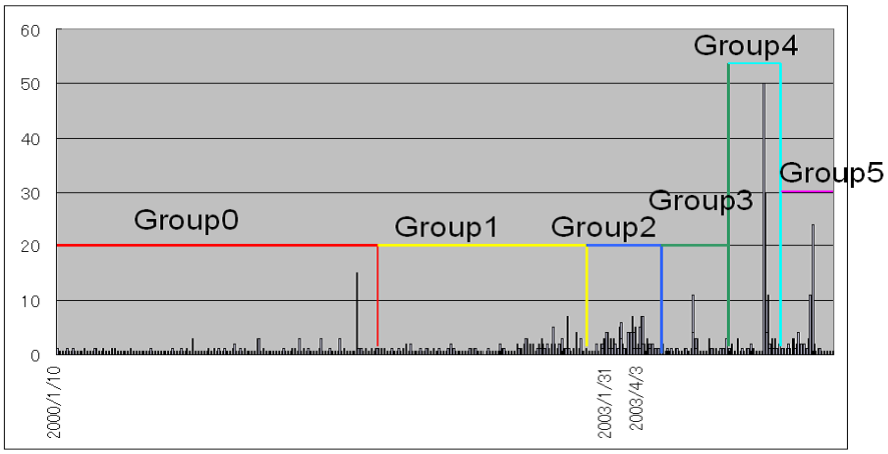

図 4.6: Hussein のクラスタリング結果

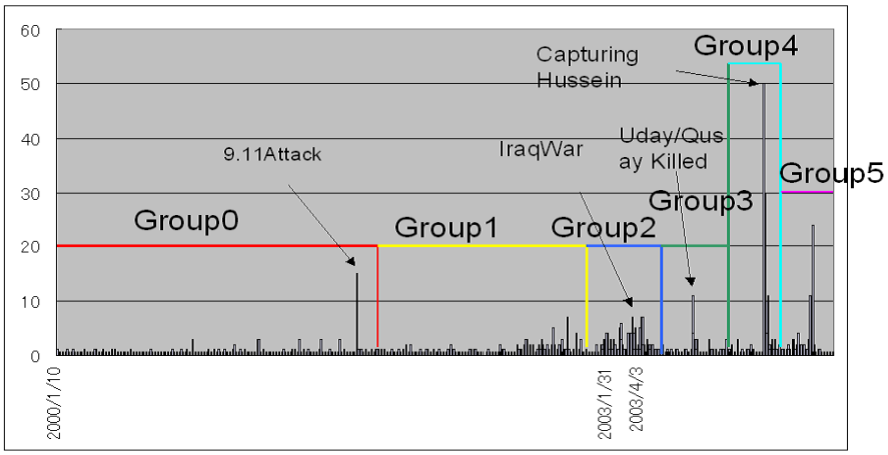

図 4.7: 実際の事件と時制クラスタの対応

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 4.7 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．

4.6.3 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 4.2 に示す．表 4.2 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．

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 4.3，それに基づくトラッキングの結果を図 4.8 に示す．

図 4.8 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．

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 4.2: サブクラスタ詳細

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 4.3: サブクラスタの類似度

4.6.4 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 4.4).

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 4.4: 抽出された語

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 4.5).

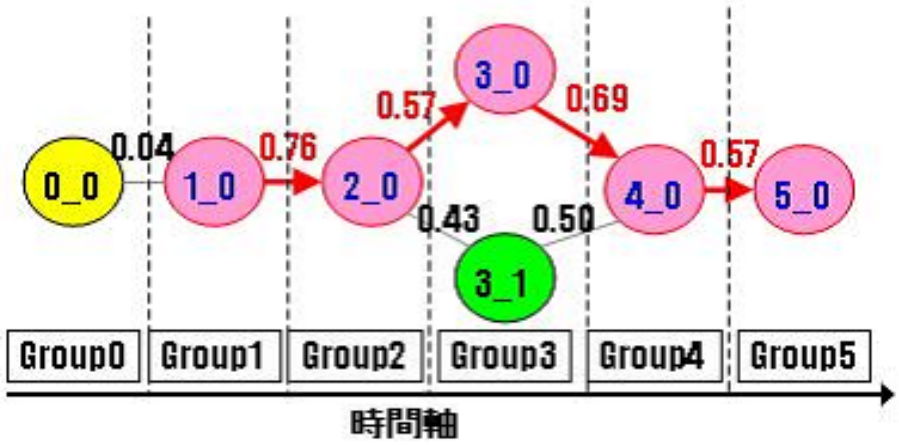

図 4.8: Tracking 結果

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

表 4.5: ラベル

4.6.5 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[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 4.8）.

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

サブクラスタ 1-0

武装: forc, militari, armi, troop, attack

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

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 はサブトピックと言える．

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る．

”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，サブトピックであると言える．

4.6.6 評価

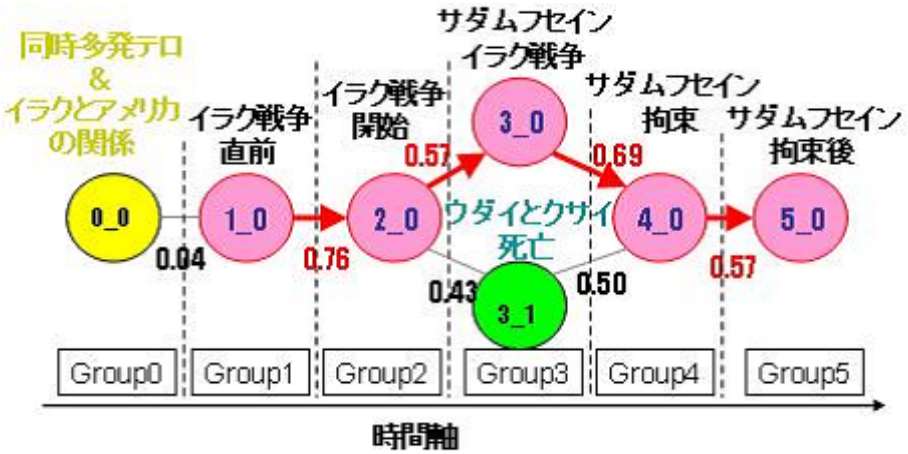

図 4.9: 追跡の結果

実験結果より，3 つのトピックが得られ（図 4.9），隣接するクラスタのサブクラスタとの関係性，他のサブトピックとの関係性を検証し 3 つのサブトピックが，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．

4.7 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．

第5章　結論

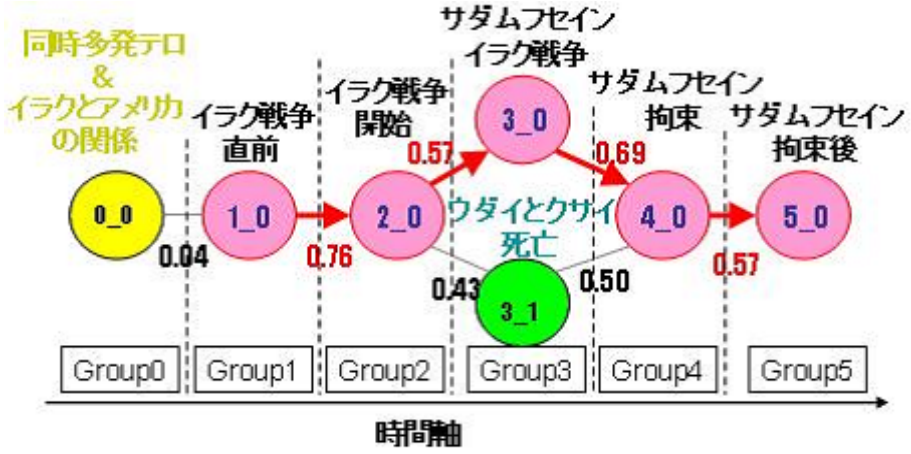

図 5.1: 追跡の結果

本研究では，検索エンジンに検索語を与えて得られた Web ページ集合から，事象の検出，追跡，要約を行う手法を論じた．実験により提案する手法を用いれば，利用者が検索エンジンから取得した Web ページ集合から容易にトピックを把握でき，利用者の検索の負担を軽減することができることを示した．

まず，Web ページからの有効時間の決定方法を 3 つの時間情報 (内容時間，作成時間，更新時間) から決定した．いくつかの実験の結果から，有効時間の決定方法を内容時間を優先的に有効時間とし，作成時間を次に有効時間とした得られた有効時間の決定方法を用いて Web ページから有効時間を抽出し，時間軸でクラスタリングを行うことで事象の検出が行えることを示し，更に，KeyGraph を用いて，Web ページの主張点を捉えた単語を抽出し，時制を考慮することで事象にラベル付けによる要約を行った．

次に，Suffix Tree を用いて語の並び，KeyGraph を用いて得られた Web ページの主張点を捉えた単語を抽出し，時制を考慮することで事象に対してラベル付けによる要約を行った．語の並びを考慮することで，単語のみのラベル付けよりも抽象度の高い高度な要約を実現した．

また，トピックの古い概念を捨て新しい概念を取り入れる追跡手法を提案した．各クラスタごとに KeyGraph を用いて基本概念を抽出し，Complete-Link クラスタリングを行いサブクラスタを構築した後，隣あう時制クラスタのサブクラスタの基本概念

の内積を計算した．そして，二つのサブクラスタの関連性を内積から判断することで，トピックの概念の変化に対応した追跡が可能となった（図 5.1）．

事象を検出するために利用した情報は，時間・時点データの推定結果である．事件にラベル付けすることは，データ集合から意味内容を抽出することであり，スキーマは名前 (記号) あるいは要約 (意味記述) として表現できる．時系列性を見出すために変化の追跡が必要である．そこで基本的な概念（基本概念）を見つけ，これらの関連を追跡することで，大域スキーマとしてのトピックが検出できることを示した．検索エンジンで時間性の強いニュース情報については，データベーススキーマ的な属性を与えることができると言える．

今後の課題として，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．

謝辞

本研究を遂行するにあたり，日頃より数々のご指導をいただいた，法政大学工学部情報電気電子工学科 三浦孝夫教授に深く御礼申し上げます.

また，産能大学経営情報学科 塩谷勇教授にも多くのご指導をいただきました．深く感謝いたします.

データ工学研究室の先輩方，同輩，後輩たちにも，本研究の遂行にあたって数多くの助言と快適な研究環境の整備をして頂きました．御礼申し上げます.

修士論文として私の研究をまとめることができたのも，多くの皆様方の御支援，御協力の賜物であります．この場をお借りしまして，厚く御礼申し上げます.

最後に，今までの学生生活を支えてくださった私の両親に感謝したいと思います.

参考文献

[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] 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] 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

[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