2004－DBS－134　(II)　(41)
2004／7／14

# 話題構造グラフを用いたニュースステーマの時間的推移抽出手法の提案

林 英俊† 手塚 太郎† 小山 聡† 田中 克己†

†京都大学情報学研究科社会情報学専攻 〒606-8501 京都市左京区吉田本町

E-mail: †{hhayashi, tezuka}@dl.kuis.kyoto-u.ac.jp, oyama@kuis.kyoto-u.ac.jp, tanaka@dl.kuis.kyoto-u.ac.jp

**あらまし** 本論文では、話題構造グラフのグラフ属性を利用して、ニュース記事集合から時間的推移を抽出するために記事間関連を量的側面と質的側面との両面から分析する。まず量的側面については、単語レベルではなく、話題構造グラフレベルでの類似度計算によって、記事間関連の強さをより正確に表現できることを示す。次に質的側面については、話題構造グラフにおける節点の次数というグラフ属性を用いて、記事間の共通キーワードを分類した後、記事間関連の意味を分析する。最終的には分析結果を用いて、具体的な語を中心としてニュースステーマの時間的推移を抽出する手法を提案する。

**キーワード** 話題構造グラフ，データマイニング，時系列データ

# A Topic Structure Graph Approach for Discovering Topic Changes of News Articles

Hidetoshi HAYASHI† Taro TEZUKA† Satoshi OYAMA† and Katsumi TANAKA†

†Department of Social Infomatics, Kyoto University Yoshida Honmachi, Sakyo-ku, Kyoto-shi, 606-8501 Japan

E-mail: †{hhayashi, tezuka}@db.soc.i.kyoto-u.ac.jp, oyama@kuis.kyoto-u.ac.jp, tanaka@i.kyoto-u.ac.jp

**Abstract** In order to extract topic changes of news themes from a set of news articles, relationships between articles are analyzed from both quantitative and qualitative aspects, based on Topic Structure Graph. First, the quantitative analysis proves that similarity calculation at the level of Topic Structure Graph can express the strength of relationships between the articles more precisely than similarity calculation at the level of keywords. Second, the qualitative analysis categorizes the common keywords between the articles based on node degree and gives the sense of relations between the articles. Finally, based on these results, we proposed a method of extracting word-centric topic changes of news themes.

**Keyword** Topic Structure Graph, Data Mining, Time-series Data

## 1. 序論

通常の文章とは異なり、ニュース記事は時間属性を持っている。ニュース記事集合は時間属性を持った記事の集合であり、その集合中には記事同士が関連しあう時間軸に沿った流れが存在するはずである。本研究は，この時間軸に沿った流れを抽出することを目標とする。記事と記事間関連の集合である時間的推移を抽出するためには、記事間関連を分析する必要がある。

従来のニュースサイトなどは、カテゴリ分類によって記事間関連を扱っているが、ニュースの内容は予測不可能なため、用意されたカテゴリでは対応しきれない、または分類できたとしても経済・社会といった大きな概念レベルにとどまる。また、時間属性を扱った従来の研究は、動画処理分野に属するものが多い。動画に関するほとんどの研究は、シーン検出・テロップ認識・音声抽出を中心としており，シーン分解後のシーン間関連の分析については、単語レベルの類似度計算手法を用いてシーン間関連の強さの量を測るにとどまっている。

本研究では時間的推移を抽出するために、記事から語のネットワークである話題構造グラフを作成し、話題構造グラフレベルで記事間関連を量的側面と質的側面の両面から解析する。まず量的側面については、単語レベルではなく、話題構造グラフレベルでの類似度計算によって、記事間関連の強さを表現する。話題構造グラフの枝を利用して類似度計算手法を行った場合、従来の単語レベルの類似度計算手法より番組作成者の意図に近づいたことを示す。次に質的側面については、話題構造グラフにおける節点の次数というグラフ属性を用いて、記事間関連の意味を考える。2 記事の共通キーワードがそれぞれの記事にとってどのような位置を占めるかを $tf*idf$ 値と節点の次数から幹・根・枝に分類する。その後、共通キーワードが占める位置の組み合わせに応じて 2 記事間関連の意味を決定する。最終的には、語に関してテーマの広がりと深まりをミクロに分析することで、ニュースステーマの時間的推移を抽出する方法を提案する。

以降、2 章では話題構造グラフとは何かを定義し、3 章では関連研究を紹介する。4 章では話題構造グラフレベルでの類似度計算によって、記事間関連の強さを扱う。5 章では話題構造グラフにおける節点の次数を利用して、記事間関連の意味に基づき分類した後、時間的推移抽出へ応用する手法を提案する。最後に 6 章で結論を述べる。

## 2. 話題構造グラフ

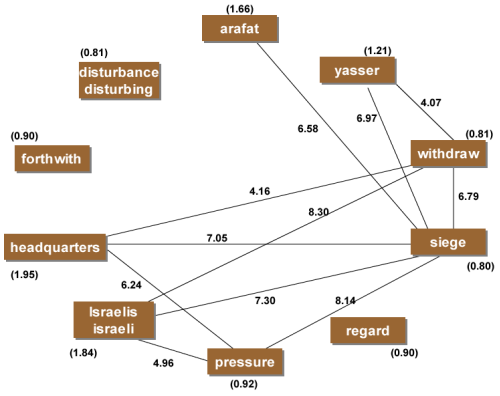


**図 1: 話題構造グラフの例**

話題構造グラフとは，図１のような重みつき無向グラフである．グラフの節点が語であり、節点の重みがその語の重要度を表す。グラフの枝は語間関連であり、枝の重みは関連の重要度を表す。単なるキーワード集合と比較して、話題構造グラフはキーワード間関連を含むという点で、記事をより文脈を含んで正確に表現できる。また、構文解析に対して、表現の精密さでは劣るが、計算時間が短いため大量の記事を扱う上で有利である。このように、話題構造グラフは手軽に文脈を考慮するためのツールとして有効である。話題構造グラフ作成方法を図2に示し以下で説明する。

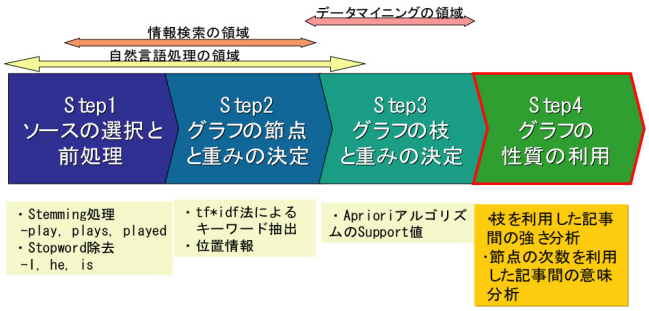


**図 2: 話題構造グラフ作成ステップ**

### 2.1. ソースの選択と前処理

自然言語テキストをソースとする。まずは、ストップワード処理を実行して，記事を特徴付ける上で役に立たないストップワード(I，yes，will) を除去する．代表的検索システム SMART システム[1]で標準的に使われているストップワードリストを使用した．次に、Stemming 処理を行って、”Israeli”と”Israel”のような意味の重複を取り除く．Porter のアルゴリズム[2]用いて，様々な語形から語幹を抽出して，語形の多様性を正規化する．

### 2.2. グラフ節点と重みの決定

このプロセスでは、*tf*idf* 法[1]を適用することによって各記事を特徴付ける重要なキーワードを抽出する．記事 $D_i$ におけるターム $t_j$ の出現頻度を $freq(i, j)$， $D_i$ 中の総ターム種類数を $Kind(i)$，ターム $t_j$ が出現する文書数を $df_j$，記事総数 $N$ として，

$$tf_{ij} = \frac{\log(freq(i,j)+1)}{\log(Kind(i))} \qquad idf_j = \log(\frac{N}{df_j})$$

*tf* 値はタームの網羅性を，*idf* 値はタームの特定性を表現している．記事 $D_i$ におけるターム $t_j$ の重みを

$$w_j^i = tf_{ij} \times idf_j \qquad (1)$$

として，この値で記事における単語の重要性を決める。ランキングされたうち単語トップ 10 が話題構造グラフの節点に，$w_j^i$ の値が節点の重みとなる。

### 2.3. グラフ枝と重みの決定

このプロセスでは相関ルール抽出の *Apriori* アルゴリズム[2]を 2.2 で抽出したキーワード群に適用することでキーワード間の隣接関係を調べる。文レベルでのキーワード共起を調べ、*seige - israel* のように，相関の左辺と右辺を１単語に限定した相関ルールを求める．ある文中にターム $t$ が出現する確率を $P(t)$ とし， 同じ文中に $t_{j_1}$ と $t_{j_2}$ が出現すれば共起と判断すると，ターム $t_{j_1}$ と $t_{j_2}$ の間の *Sup* 値と *Conf* 値は下式のように定義される．

$$Sup(j_1, j_2) = P(t_{j_1} \cap t_{j_2}) \qquad Conf(j_1 \Rightarrow j_2) = P(t_{j_2} \mid t_{j_1})$$

このうち、*Sup* 値を期待値 0，分散 1 に正規化した値を、単語間の共起度(*Ass*)と定義する。記事 $D_i$ 中において、*Sup* 値の平均を $\mu$、分散を $\sigma^2$ とすると、ターム $t_{j_1}$ と $t_{j_2}$ 間の共起度( $Ass^i(j_1, j_2)$ )は(2)式のようになる．

$$Ass^i(j_1, j_2) = \frac{Sup(t_{j_1}, t_{j_2}) - \mu}{\sigma} \qquad (2)$$

共起度が 0(閾値)以上の関係が話題構造グラフの枝に、共起度の値が枝の重みとなり、話題構造グラフが完成する。

### 2.4. グラフの性質の利用

話題構造グラフは直感的には単語と隣接関係を用いた、記事の簡単な要約を表している。前述の図１は、アラファト議長包囲解除の記事の話題構造グラフである。本研究ではこの話題構造グラフにおけるグラフ属性と記事における性質との対応関係を明らかにした後、グラフ属性を利用した計算手法を考える。4 章では枝というグラフ属性を用いて、記事における隣接関係を考慮した類似度計算手法を提案する。5 章では、節点の次数というグラフ属性を用いて、記事における重要部分網羅度を考慮したキーワード分類手法を提案する。

## 3. 関連研究

### 3.1. テキストマイニング

テキストから単語間の関連などの知識を抽出する研究はテキストマイニングと呼ばれる。Hearst[3]は、テキストマイニングとは構造化されていないテキストデータから新知識を発見する技術であり，まず自然言語処理がに加えて、情報検索，データマイニングや機械学習の技術を組み合わせて実現されると定義している．

Rajman[4]らは各単語の品詞を調べることで語をターム（例えば credit card）単位で扱う．さらに不要品詞

[1] Salton, G., Ed, ”The SMART Retrieval System"
[2] M. F. Porter, “An algorithm for suffix stripping”

の除去，接辞処理などの前処理を行った後，相互情報量などの統計的手法を用いて重要タームを決定する．その後、重要ターム間で相関ルール抽出を行っている。

これらの研究は単語間関連に焦点を当てているのに対し、我々の研究の焦点は単語間間関連に着目したグラフの枝の選定ではなく、グラフ属性の利用にある。

### 3.2. 視覚化に関する研究

情報検索やマイニングの結果を視覚化する研究は多い．WebBrain[3]や Cat-a-Cone Interface[4] は，カテゴリ関係を可視化して Web ナビゲーションに利用している．

渡部[5]はスプリング埋め込み技術を用いてテキストマイニングによって導かれた連想関係を可視化している。TouchGraph[5]は、情報の関連を視覚化して動的なナビゲーションシステムを構築するためのプログラムを提供している。 これを利用して、AmazonBrowser[6]は商品と同時購買される商品間関連を、GoogleBrowser[7]は Web ページとその間のリンクを視覚化したシステムを構築している。

ナビゲーションが最終目的である視覚化の研究とは異なり、我々はグラフ属性を活用して従来にはない計算手法を実行するために話題構造グラフを作成する。

### 3.3. テーマの時間的推移を捉える研究

TDT[6]は，配信ニュースデータを各ニュースに分割して,その分割されたニュースをトピックごとに分類する。さらに、新たに配信されたニュース記事も、既存トピックとの関連を考慮して分類し、ニュースデータを利用者にわかりやすい形で提供している．

角谷ら[7]は、過去の記事との類似度＋配信時間の差に加えて、続報予定という概念を用いて続報リストを作成するという、マルチチャンネル型配信システムのための時間情報に基づくクラスタリング方式を提案している。

これらの時間属性を生かしたトピック推移に関する研究が記事間関連の強さを単語レベルの類似度計算で考慮するのに対して、本研究は話題構造グラフレベルでその記事間関連の質まで考慮することが相違点である。

### 3.4. グラフ構造を利用した研究

高橋ら[8]は構文解析技術を利用して、自然言語テキストから重みなしの構文木を作成した後、構文木間の類似度を、"構文木間の内積=共通する部分木の数"と部分儀の構造を用いて定義している。庄田ら[9]は全連結部分グラフの節点の次数を次元としたグラフスペクトルによってグラフ構造を定量化した後、無向グラフ間類似度をグラフスペクトル間の類似性と定義している。

松尾ら[10]は SmallWorld というグラフ構造を利用して、著者の主張を伝える上で重要な語を記事から抽出している。SmallWorld とは、節点がクラスタ化されているにもかかわらず任意の 2 点間のパス長が短いグラフのことであり、このグラフ構造においては特定の節点がパス長を縮めるのに大きく貢献している。記事から得た語のネットワークが SmallWorld 構造を持ち、その中で特定の節点となる語が元の文書において著者の主張語であるということを評価実験を通じて示している。大澤ら[11]は、語からなるネットワークを土台・柱・屋根の 3 種類の部分に分けてグラフを作成し、文書の主張を表すキーワードを抽出する手法を提案している。土台は文書が元にしている基本概念を表し、文書中における頻度によって求める。屋根が著者の主張点を表し、土台単語との共起から求める。柱は内容の主な展開を表しており、屋根-土台間の関連を調べることによって求める。

Ma ら[12]は、主題語と内容語から構成されるトピックストラクチャの補完度によってトピック間関連について調べている。トピックストラクチャ同士を結合させたときの、幅の増加を情報網羅度の広がりに、深さの増加を情報詳細度の深まりに対応させて、トピック間のつながりを分析している。灘本ら[13]は、各記事をメイントピックとサブトピックからなるトピックグラフによって表現する。比較元の文書のメイントピックと類似し、かつ比較元のサブトピックと相違するトピックグラフを持つ文書をコンテクスチュアルページと定義して抽出することで、トピックグラフ間の時系列変化を抽出している。

[8][9]は類似度という関連の強さの量に焦点を当てており、[10][11]はグラフ構造を用いて各文書内の単語に焦点を当てている。本研究は、グラフ構造を用いて関連の強さに加えて意味を考え、単語ではなく単語間関連・記事間関連に焦点を当てている。[12][13]は本研究に近い研究だが、本研究が文単位での共起というミクロな位置情報を利用して、情報の広がりや深まりをトピック中心ではなく単語中心に扱うという点で異なる。

## 4. 話題構造グラフによる記事間関連の強さ分析

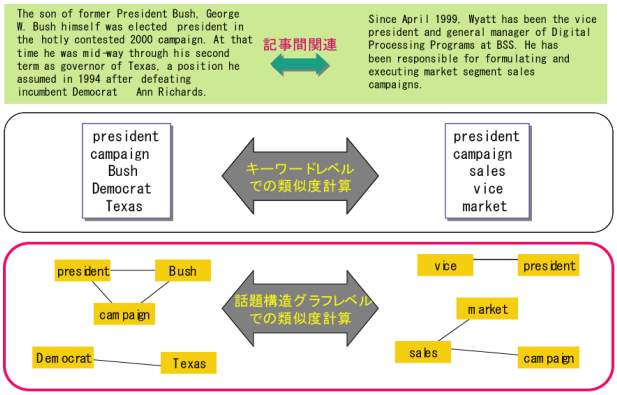


**図 3: 話題構造グラフを用いた類似度計算手法**

本章では、ニュース記事間関連の強さ（量的側面）を話題構造グラフレベルで解析する。記事間関連の強さを考えるとき、各記事をキーワードからなる特徴ベクトル間のコサイン類似度を求める手法がよく用いられる．この従来の類似度計算手法は，各記事が重要単語集合で表現できるという前提に立っているが，実際の記事は単語以外にも，共起，係り受け，段落といった文脈情報を含

[3] WebBrain http://www.webbrain.com
[4] Cat-a-Cone Interface http://www.sims.berkeley.edu/~hearst/cac-overview.html
[5] TouchGraph http://www.touchgraph.com
[6] AmazonBrowser http://www.touchgraph.com/TGAmazonBrowser.html
[7] GoogleBrowser http://www.touchgraph.com/TGGoogleBrowser.html

んでおり，それらが類似度計算に反映されないという問題がある．そこで本研究は、図３のように各記事を単語集合ではなく、話題構造グラフで表現する。その後、話題構造グラフレベルでの類似度計算を利用して、記事間関連の強さを表現する。

### 4.1．枝と元の記事の対応

2.3.のように定義したので、話題構造グラフの枝は元の記事で枝の両端の節点を表す単語の隣接関係（共起関係）を示している。そのため、枝属性を利用することは、隣接関係という文脈に関する特徴を考慮することになる。

### 4.2. 話題構造グラフによる類似度計算アルゴリズム

#### 4.2.1. 従来のキーワードによる類似度計算

記事 $D_i$ に出現する単語の *tf*idf* 値を成分とする特徴ベクトルを $w^i=(w_1^i,w_2^i,.....,w_n^i)$ とする．記事 $D_{i_1}$ と $D_{i_2}$ の *tf*idf* 値に関する特徴ベクトルをそれぞれ $w_1^i$，$w_2^i$ とすると，記事 $D_{i_1}$ と $D_{i_2}$ の類似度 $Sim(i_1,i_2)$ は下式で表される．

$$Sim(i_1,i_2)=\frac{w_1^i\cdot w_2^i}{\sqrt{\left|w_1^i\right|\left|w_2^i\right|}} \qquad (3)$$

#### 4.2.2．話題構造グラフによる類似度計算

4.1.1.で考えた *tf*idf* 値を成分とする特徴ベクトルに加え，単語間関連を示す共起度(*Ass*)を成分とする特徴ベクトルについても考える．記事 $D_i$ に出現する単語の *tf*idf* 値を成分とする特徴ベクトルを $w^i=(w_1^i,w_2^i,.....,w_n^i)$，単語間の *Ass* 値を各成分とする特徴ベクトルを $a^i=(a_{11}^i,a_{12}^i,a_{13}^i,.....a_{n-2,n-1}^i,a_{n-1,n}^i)$ とする．記事 $D_{i_1}$ と記事 $D_{i_2}$ それぞれの *tf*idf* 値に関する特徴ベクトルを $w_1^i$，$w_2^i$，*Ass* 値に関する特徴ベクトルを $a_1^i,a_2^i$ とすると，記事 $D_{i_1}$ と $D_{i_2}$ の類似度 $Sim(i_1,i_2)$ は下式で表される．$\alpha$，$\beta$ は２種類の類似度を重み付けのためのパラメータである．

$$Sim(i_1,i_2)=\alpha\frac{w_1^i\cdot w_2^i}{\sqrt{\left|w_1^i\right|\left|w_2^i\right|}}+\beta\frac{a_1^i\cdot a_2^i}{\sqrt{\left|a_1^i\right|\left|a_2^i\right|}} \qquad (4)$$

### 4.3. 評価実験

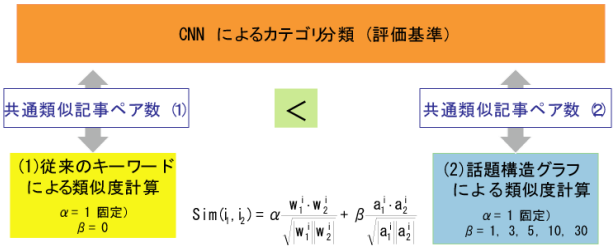


**図 4: 評価実験の概要**

本節では，話題構造グラフによる類似度計算手法が従来の類似度計算手法に比べて，記事作成者のカテゴリ分類結果との距離という基準で正確であることを示す．

CNN の音声情報テキスト TRANSCRIPT10 日分を実験データとして使用した．類似度計算の正確性を評価するための評価基準としては CNN が記事分類に用いるカテゴリ分類キーワードを採用した．このカテゴリ分類キーワードは，ニュース製作者が手動で作成して各記事に付与している．我々は人間が類似していると考える結果に近づけることを目標とするため，記事作成者のカテゴリ分類結果との距離を類似度計算の正確性として定義した．

図 4 に示した評価実験の手順について以下で説明する．まず CNN によって与えられたカテゴリ分類キーワードを元に，類似記事ペアトップ $N$ を算出し，定量的評価実験の評価基準とする．次に，$\alpha=1$ と固定して，従来のキーワードによる手法 $\beta=0$ と話題構造グラフによる手法 $\beta=1,3,5,10,30$ の２通りの手法で類似記事ペアトップ $N$ を算出する．最後に，それぞれの類似記事ペアトップ $N$ と評価基準の類似記事ペアトップ $N$ との間の共通類似記事ペア数によって２手法の正確性を測定する。

**(1) $\beta$ 別共通類似記事ペア数（$\alpha=1, N=20$）**

| | β=0 | β=1 | β=3 | β=5 | β=10 | β=30 |
|---|---|---|---|---|---|---|
| 2002/9/29 | 10 | 11 | 11 | 11 | 11 | 11 |
| 2002/10/6 | 18 | 19 | 19 | 19 | 19 | 19 |
| 2002/10/13 | 2 | 3 | 3 | 3 | 3 | 3 |
| 2002/11/3 | 5 | 5 | 5 | 5 | 5 | 5 |
| 2002/11/10 | 10 | 11 | 11 | 11 | 11 | 11 |
| 2002/11/17 | 5 | 6 | 6 | 6 | 6 | 6 |
| 2002/12/15 | 2 | 3 | 4 | 4 | 4 | 3 |
| 2002/12/22 | 12 | 11 | 13 | 13 | 12 | 12 |
| 2003/6/1 | 8 | 9 | 9 | 8 | 7 | 7 |
| 2003/6/8 | 13 | 13 | 13 | 13 | 13 | 13 |
| 合計 | 85 | 91 | 94 | 93 | 91 | 90 |
| 向上point(β=0の合計を100) | 0 | +7.1% | +10.6% | +9.4% | +7.1% | +5.9% |

ノードの重み重視　　エッジの重み重視

**表 1: $\beta$ 別共通類似記事ペア数（$\alpha=1, N=20$）**

様々な $\beta$ 値に対する比較性能が向上度を，$\beta=0$ の合計値を 100 としたときの向上ポイント(%)で表す．表１は $\alpha=1$ と固定して，従来のキーワードによる手法( $\beta=0$ )と話題構造グラフによる手法( $\beta=1,3,5,10,30$ )での実験結果である．この実験では $N=20$ と固定して，評価基準との共通類似記事ペア数トップ 20 を求めた．従来の手法と比べ，話題構造グラフによる手法は全 $\beta$ 値に対して比較性能が向上しており，$\beta=3$ で最大 10.6%をとっている．

**(2)トップ $N$ 別向上ポイント推移（$\alpha=1$）**

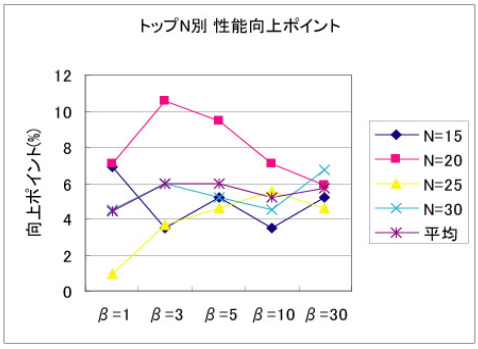


**表 2: トップ $N$ 別向上ポイント推移（$\alpha=1$）**

(1)の結果が $N$ に依存していないことを示すために、予備実験として様々な $N$ をとった場合の比較性能の向上ポイントを調べる．表 2 は $\alpha=1$ と固定して，話題構造グラフによる手法( $\beta=1,3,5,10,30$ )による，(1) で定義した向上ポイント(%) の推移を表示した結果である．平均して 5%前後の向上ポイントを取っており、$N$ に依存せずに比較性能が向上していることがわかる。

## 5. 話題構造グラフによる記事間関連の意味分析

本章では、記事間関連の意味（質的側面）を話題構造グ

ラフレベルで解析する。記事間関連を扱う場合、従来は類似度計算という統計的指標で強さを表すことが多く、そのつながりが持つ意味についてはあまり議論されてこなかった。本研究では、節点の次数というグラフ属性を用いて、記事間関連の意味を分類する手法を提案する。

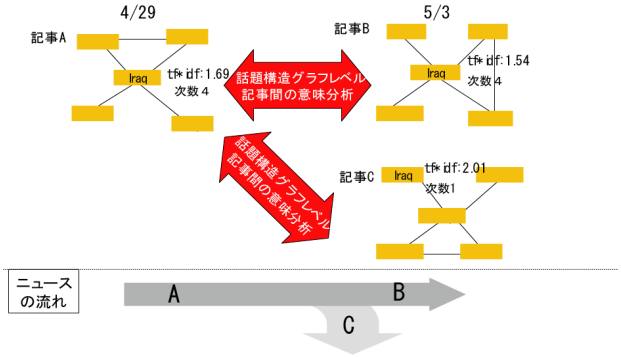


図 5: 話題構造グラフレベルでの記事間関連の意味分析

## 5.1. 節点の次数と元の記事との対応

節点の次数とはその節点から出ている枝の本数である。話題構造グラフにおいて次数の大きい節点は、元の記事において多種類のキーワードと共起しているキーワードを意味している。記事においてキーワードが表れる部分が重要であるとすると、このような節点は記事における重要部分を網羅している。つまり、話題構造グラフにおける節点の次数は、元の記事において重要部分網羅度に対応する。よって、記事 $D_i$ におけるターム $t_j$ の網羅度 $\mathrm{deg}ree_j^i$ を以下のように定義する。

$$\mathrm{deg}ree_j^i = (\text{記事}D_i\text{の概念グラフにおけるノード}j\text{の次数}) \quad (5)$$

## 5.2. *tf*idf* 値と *degree* 値に基づくキーワード分類

*tf*idf* 値は語の統計的重要度を表しており、記事中における位置情報までは含んでいない。*degree* 値は重要部分をどれだけ網羅するかという記事中の位置情報を踏まえた値である。この異なる意味を持つ 2 値を軸としたキーワード分類を考える。キーワードがそれぞれ相対的に記事中でどのような役割を持つかを 4 種類に分類する。

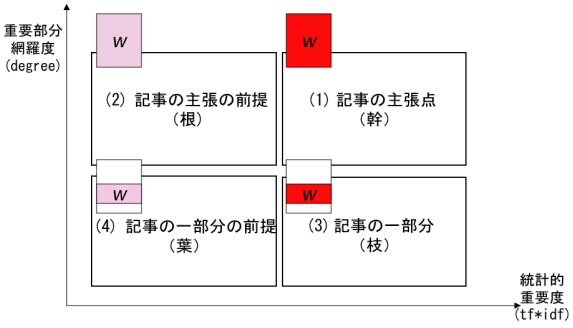


図 6: *tf*idf* 値×*degree* 値によるキーワード分類

- **(1)幹 …記事の主張点**
  記事を統計的に特徴付け、かつ重要部分を網羅しているため、記事の主張が表れる語である。
- **(2)根…主張の前提**
  強調はされないが記事の重要部分を網羅している。そのため、この記事においては強調されないが以前に説明されたような、主張が前提とする語である。
- **(3)枝…主張の部分的根拠**
  強調されるが重要部分の一部に偏って現れる。そのため、一部分で主張を集中的に裏付けるような、主張の部分的根拠となる語である。

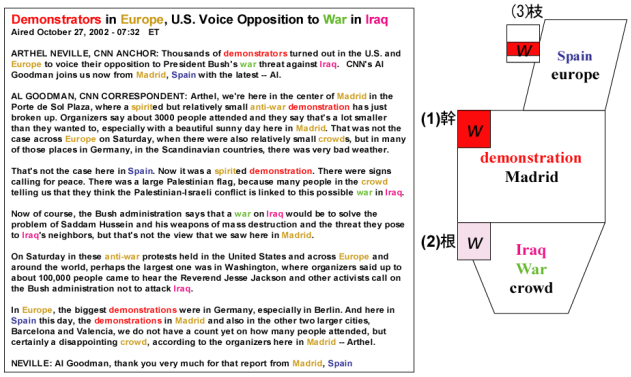


図 7: イラク反戦記事におけるキーワード分類実験

CNN のイラク反戦記事(2002/10/27)に本分類手法を適応した実験結果を図 7 に示す。まず”demonstration”は *tf*idf/degree*=2.79/8 で、両値とも高く幹に分類される。”demonstration”は記事中に分散してよく表れ、記事において著者の主張を表す語となっている。”Iraq”は 1.27/6 で、*tf*idf* が低い割には *degree* 値が高く根に分類される。この記事は Iraq での開戦ムードを前提として反戦運動について述べているため、”Iraq”は記事において前提を表す語である。“Spain”は 1.94/3 で、*tf*idf* 値が高い割に *degree* 値が低く枝に分類される。”Spain”は記事中では事件の起こった”Madrid”の近くに限って現れ、特定部分を付帯的に説明する語である。

## 5.3. 記事間関連の定性的分析

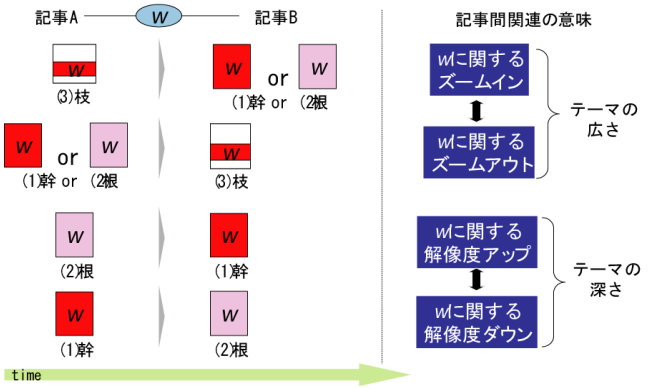


図 8: 記事間関連の意味の相違

本研究では図 7 右部分のように、ニュースのテーマを幹・根・枝となる語の集合であると定義する。本節では記事間関連が持つ意味、特にテーマの一部分を構成する語に関してのテーマの広がりと深まりをミクロに調べる。

5.2.の分類手法に従って、2 記事をつなぐ共通キーワード $w$ がそれぞれの記事にとってどのような役割を持つかを考えて、その役割の組み合わせに応じて記事間関連の意味を見出す。2 記事 A,B（time(A) < time(B)）が共通キ

ーワード $w$ でつながっている場合の記事間関連の意味を、カメラレンズの機能にたとえて図８のように定義する。“(3)枝⇒(1)幹，(2)根”という変化は記事Aでは特定部分であった $w$ が記事Bでは全体を網羅しており、$w$ に焦点を合わせて視野を狭めているため、**$w$ に関するズームイン**と定義する。逆に、“(1)幹，(2)根⇒(3)枝”は $w$ から視野を広めているため、**$w$ に関するズームアウト**と定義する。これら２つはテーマの広がりに関する推移である。“(2)根⇒(1)幹”は記事Aでは前提であった $w$ がBの主張点なっており、$w$ に関して詳細に見ているため、**$w$ に関する解像度アップ**と定義する。逆に“(1)幹⇒(2)根”は $w$ に関する詳細度を落としているため、**$w$ に関する解像度ダウン**と定義する。これら２つはテーマの深さに関する推移である。最後に同じものへの変化“(1)幹⇒(1)枝，(2)根⇒(2)根”はこれまでの流れを引き継いでおり、**$w$ に関する続報**と定義する。

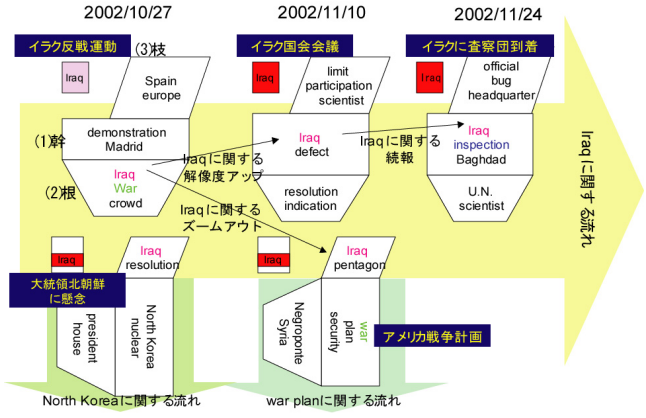


**図 9: 時間的推移抽出実験(2002/10/27～2002/11/24)**

CNN ニュース(2002/10/27～2002/11/24)の記事集合を対象に、図８の分類基づいて Iraq に関する時間的推移を抽出実験した結果を図９に示す。キーワード Iraq が記事に占める役割に着目して、2 つの異質な流れを検討する。

イラク反戦運動(10/27)⇒イラク国会会議(11/10)⇒イラクへ査察団到着(11/24)は Iraq に関する解像度アップ＋続報であり、まず反戦運動の記事からイラク国会会議へと Iraq に関して深まり、次にイラクと国連を主体としたイラク大量破壊兵器関連のせめぎあいに関する記事へと続く流れを示している。これに対して、イラク反戦運動(10/27)⇒アメリカ戦争計画(11/10)は Iraq に関するズームアウトに相当する。視点がイラクのみからアメリカも含む視点へと広がり、主体が戦争へと移る。アメリカ戦争計画(2002/11/10)はアメリカの戦争プランが主題であり、アメリカとして軍事資源の十分性やテロ対策を議論している。この記事はイラクではなく戦争に対するアメリカの国会での決議に関する流れに近づいており、大量破壊兵器関連に至る流れとは異質な流れである。

## 6. 結論

本研究では、まず、話題構造グラフを用いて記事間関連の強さを分析した。グラフの枝を利用する類似度計算手法を提案し、評価実験を通じてニュース作成者の意図に近い結果が出ることを示した。次に、話題構造グラフを用いて記事間関連の意味を分析した。節点の次数を利用して共通キーワードをそれぞれの記事において占める役割に応じて分類した後、記事間関連の意味の違いを決定した。最後に、語に関してのテーマの広まりと深まりを分析し、単語中心にテーマの時間的推移を抽出する手法を提案した。今後も、ニュース集合を対象に時間的推移を抽出する実験を継続し、語レベルでのテーマの推移、語の寿命などを調べる。その過程で、記事間関連の意味の相違について厳密な定式化を行った後、その妥当性を検証するための評価実験を行うことが今後の課題である。


## 謝辞

本研究の一部は，文部科学省科学研究費特定領域研究(2)の「Webの意味構造発見に基づく新しいWeb検索サービス方式に関する研究」（課題番号：16016247）による．ここに記して謝意を表します．


## 文　献


[1] G. Salton and C.S. Yang, “On the Specification of Term Values in Automatic Indexing”, Journal of Documentation 29(4), pp351-372 December 1973

[2] R. Agrawal, R. Srikant: "Fast Algorithms for Mining Association Rules", *Proc. of the 20th Int'l Conference on VLDB*, Santiago, Chile, Sept. 1994.

[3] Marti A. Hearst, “Untangling Text Data Mining”, ACL ’99, pp.03-10, 1999

[4] Feldman R., Fresco M., Yakkov K., Lindell Y., Liphstat O., Rajman M., Schler Y., and Zamir O., “Text Mining at the Term Level”, PKDD ’98, 1998

[5] 渡部勇，“ビジュアルテキストマイニング”，人工知能学会誌，Vol.16-2，pp.226-232，2001.

[6] Allan, J., Carbonell, J.G., Doddington, G., Yamron, J. and Yang Y., “Topic Detection and Tracking Pilot Study Final Report.”, In Proceedings of the Broadcast News Transcription and Understranding, Feb. 1998.

[7] 角谷和俊，松本好市，高橋美乃梨，上原邦昭，“マルチチャンネル型ニュース配信システムのための時系列クラスタリング”，情報処理学会論文誌，Vol.43，No.TOD14，pp.87-97 ，2002.

[8] 高橋哲朗，乾健太郎，松本裕治，“テキストの構文的類似度の評価方法について”，情報処理学会自然言語処理研究会, NL-150-24, Jul. 2002.

[9] 庄田良介，松田喬，吉田哲也，元田浩，鷲尾隆，“構造的類似性に基づくグラフクラスタリング”，JSAI2003.

[10] 松尾豊，大澤幸生，石塚満，“Small World 構造を用いた文書からのキーワード抽出”，43-6， 情報処理学会論文誌，June 2002.

[11] 大澤幸生，Benson，N.E.，谷内田正彦，“KeyGraph: 単語共起グラフの分割統合によるキーワード抽出”，J82-D1-2，電子情報通信学会論文誌，Feb. 1999.

[12] Qiang Ma, Katsumi Tanaka, “Topic-Structure Based Complementary Information Retrieval for Information Augmentation”, Proceedings of the 6th Asia Pacific Web Conference, Hangzhou, China, April 2004

[13] 灘本明代，田中克己，“T-CNB: 時間を考慮した文脈に基づくニュースブラウザの提案”，電子情報通信学会第 15 回ワークショップ DEWS2004, March 2004