

複数国の新聞からの多観点比較による分析 ～GDELT データを用いた分析～

Multifacet comparative analysis of newspaper articles from different countries - Analysis based on Global Database of Events, Language and Tone (GDELT) -

吉岡真治^{1*} 神門典子²
Masaharu Yoshioka¹ Noriko Kando²

¹ 北海道大学大学院情報科学研究科

¹ Graduate School of Information Science and Technology, Hokkaido University

² 国立情報学研究所

² National Institute of Informatics

Abstract: The News Site Contrast (NSContrast) system analyzes multiple news sites based on the concept of contrast set mining and it can extract terms that characterize different topics of interest for specific countries. In this study, we used the NSContrast system to analyze Global Database of Events, Language and Tone (GDELT) data by comparing news articles from different regions (e.g., USA, Asia, and the Middle East). We also present examples of analyses performed using this system.

1 はじめに

近年、世界中の報道機関が、各々の記事を配信するニュースサイトを構築し、記事を提供している。その結果、これまでは、簡単には読むことの出来なかった世界中のニュース記事を閲覧することができるようになってきた。我々は、これまでに、複数の複数の国のニュースサイト発信する記事を比較することにより、それぞれのサイトが代表する各国の興味の違いなどを分析する新聞比較システム NSContrast の開発を行ってきた [1, 2]。このようなシステムを構築する上での問題の一つは、大規模なニュースサイトから記事を集め、分析を行う必要があるという点にあった。

この様なニュースの分析をするというニーズに答えるための基盤データとして、Global Database of Events, Language and Tone (GDELT) [3]¹が、2013 年から公開されている。このデータベースは、Google News に提供されている様々なニュースサイトのデータなどを用いて構築された非常に大規模なデータベースで、世界中で報道されている毎日の記事において、どのようなイベントが、どのようなトーン（賛否）で報道されている

のかをまとめたデータベースである。

本論文では、NSContrast と GDELT についての簡単な紹介を行ったあと、NSContrast により GDELT データを分析した事例について紹介する。

2 NSContrast における GDELT データの利用

2.1 NSContrast

NSContrast[1] は、日本語で記述されている各国のニュースサイトとそれぞれの国の言語のニュースサイト（韓国：朝鮮日報、中国：新華社）から獲得した記事を機械翻訳したものをニュース記事データとして利用し、各国の興味の違いなどの分析を行うシステムである。

本システムでは、各記事データから、ChaSen による解析結果で得られた各記事に含まれる単語の一覧、CaboCha による固有名詞抽出の結果得られた人名・地名・組織名と Wikipedia のエントリーとの比較結果による Wikipedia 項目、意見分析システム [4] による、賛成、反対の意見情報が付加されている。

NSContrast では、この新聞記事データベースに対して、以下の機能を提供している。

*連絡先：北海道大学大学院情報科学研究科
〒064-0806 札幌市北区北 14 条西 9 丁目
E-mail: yoshioka@ist.hokudai.ac.jp

¹<http://gdeltproject.org/>

1. 国別注目単語比較

日中韓米の各国の新聞記事を対象に、バースト解析 [5] を行うことにより、各国ごとに、注目されている特徴的なキーワードと注目期間を表示する。

2. 今日の話題

毎日の記事をクラスタリングすることにより、その日の新聞記事で多く報道されているトピックに相当する記事群を抽出する。

3. 記事検索

単純に、キーワードを含むかどうかという検索に加え、記事に付加されている人名・地名・組織名・Wikipedia のエントリーを含むかどうかという形の問い合わせが可能。

4. 関連語の関係性分析

与えられたキーワードを含む記事群に特徴的に存在する関連語 (共起語) とその関係を可視化する。可視化の対象としては、記事中の語句をそのまま使うだけでなく、記事に付加されている人名・地名・組織名・Wikipedia のエントリーに限定して表示させることも可能である。

本システムにおいて、特徴的に存在する共起語とは、各国の記事において、記事群に対して相関性の高い共起語だけでなく、他の国と比べて相関性の変化が大きな語も候補とする。また、これらの共起語の内、さらに、ダイス係数によって、一定以上の共起性を持っている共起語の間にリンクを設定し、共起語グラフとして可視化する。

5. 関連語の時間遷移分析

与えられたキーワードを含む記事群を 3 日単位で分割し、各々の記事群に対して、特徴的に存在する関連語 (共起語) を抽出し表示する。

本機能において、特徴的に存在する共起語とは、特定の期間の記事において、記事群に対して相関性の高い共起語だけでなく、他の期間と比べて相関性の変化が大きな語も候補とする。

6. 国別比較

入力したキーワードを含む記事数の各国ごとの遷移がグラフで表示されると共に、バースト期間を可視化する。また、対照分析の結果として、関連語の関係性分析の可視化結果のグラフを表示する。

7. 多観点分析 (マルチファセット分析)

特定の検索語を含む記事に含まれるデータをいろいろな観点 (ファセット) から可視化する。本システムでは、以下の観点が利用可能である。

- キーワード

- Wikipedia のエントリー語

- 人名
- 組織名
- 地名
- 賛否

また、各観点ごとのデータは、以下の様な形でまとめて表示する。

- 時系列グラフ
- 円グラフ
- 棒グラフ
- (重みつき) 頻度の表

ユーザは、複数存在するグラフ表示領域毎に、表示させたい観点、表示領域に対して与える絞り込み条件や種類を設定することが可能であり、以下のような分析を支援する。

- 複数の観点を組み合わせた検索条件の設定
複数の観点を組み合わせた検索条件を設定すると共に、その内容を可視化する。
- 複数の国の比較
比較したい国に対応する形で複数のグラフを表示する。表示内容 (例えば、人名) などは、全体で固定し、各グラフに対して、国名を絞り込み条件として与えると、国ごとの違いを並べて見る事が可能となる。

2.2 Global Database of Events, Language and Tone (GDELT)

Global Database of Events, Language and Tone (GDELT) [3] は、Google News、BBC Newswire などから獲得した様々なニュースサイトの記事データから構築された非常に大規模なデータベースで、以下の 2 つのデータ形式でのデータ配布が行われている。

● GDELT Event Database

GDELT の基本データベースで、世界中で起こった様々なイベントに関するデータベースである。イベントは、同一性の判定が行われてイベント ID が付与される。各イベントについては、イベントの起きた日時、同一のイベントを報告している記事の件数やその記事の URL、報道されている期間などの情報が記録されている。さらに、詳細な情報として、Conflict and Mediation Event Observation (CAMEO) コード²に基づいたイベ

²<http://eventdata.parusanalytics.com/data.dir/cameo.html>

ントの種類(紛争、協力…), 関連する人物・組織の情報(国、宗教、民族、名前…)の情報や、地理情報(地名や正規化した地名に対応する地名コード)、トーン(賛否)の数値(正の値が賛成、負の値が反対を示し、その絶対値が、賛成や反対の度合いを示す)の平均などの情報が付与されている。2013年3月以前については、1ヶ月単位や1年単位でのデータが作成されており、2013年4月1日以降は、1日単位でデータの作成が行われている。

- **GDELT Global Knowledge Graph (GKG)**
GDELT Event Databaseにおけるイベントに関するより詳細なデータベースで、主に、記事を単位として作成されており、ほぼ同じ内容を述べている記事数、記事中で述べられているイベントID、記事の分類コード、記事中に含まれる地名、人名、組織名などのリスト、トーン(賛否)の数値(正の値が賛成、負の値が反対を示し、その絶対値が、賛成や反対の度合いを示す)、記事のURLの情報が付与されている。このデータについては、2013年4月1日以降は、1日単位でデータの作成が行われている。

現在、これらのデータには、GDELTの公式サイトやGoogle BigQueryなどを通じてアクセスすることが可能である。

2.3 GDELT データの利用

NSContrastにおいては、記事を単位としたデータの分析を行うため、GDELTのGDELT Global Knowledge Graph (GKG)を用いた分析を行うこととした。

多少のデータフォーマットの違いはあるものの、基本的なデータ項目の対応がとれていることから、次のような基準でデータの変換を行った。

日付 GDELT GKG が提供している日付をそのまま利用

人名、組織名、地名 GDELT GKG が提供しているデータをそのまま利用

賛否 NSContrastでは、賛否を賛成、反対、中立という3つのレベルで分析している。これに対しGKGにおいてトーン(賛否)の数値は、-100(完全に反対)から100(完全に賛成)の値をとる。よって、この値を離散化して、3つの値に分類する必要がある。本論文では、いくつかの記事データを見ながら、実験的に、-1より小さい場合を否定、-1~1を中立、1より大きい場合を肯定と分類することとした。

サイト名 サイト名はURLの情報から抽出する。ただし、複数のサイトが同一の内容を報道していた場合には、サイトの数に対応する形で、記事のデータを複製し、データとして格納する。

サイトの所属する国 NSContrastでは、国ごとの報道の違いを分析するため、各ニュースサイトを所属する国に対応づける必要がある。しかし、GDELTで扱っているサイト数の数は膨大であり、手作業で国名を割り振ることは困難であった。また、ドメイン名を利用する方法についても検討したが、多くの国で、.comや.netといったドメインが用いられており、ドメイン名にのみ依拠するのも問題であると考えられた。そのため、次のような手順で、国名の推定を行った。

1. 各国のニュースサイトの一覧を提供しているサイトである world-newspapers.com³を利用し、そのサイトの分類を利用して、サイト名とその所属する国のデータベースを作成した。例外として、BBC Newswireについては、このリストとは別に、イギリスと判定することとした。このデータベースに存在するサイトについては、このデータを利用して所属する国名を付与する。
2. 上記のデータベースに含まれていないサイトの内、国別ドメイン(.jp,.ukなど)をトップドメインに持つ場合には、そのドメイン名に対応する国名を付与する。
3. 上記の手段で判定できない場合は、所属する国名の情報を空白とする。

サイトの所属する地域 国名では、比較分析を行う単位が小さいと考えられたため、全ての国を、アメリカ合衆国、北アメリカ(アメリカ合衆国を除く)、南アメリカ、アジア、ヨーロッパ、中東、アフリカ、オセアニア、アフリカの8つの地域に分類し、国名が付与された場合には、対応する地域名を付与した。

URL GDELT GKG が提供しているURLをそのまま利用する。ただし、同一サイトに複数の記事が存在するデータについては、そのほとんどが、CGIパラメータのみが違い、最終的に同じ内容の記事が表示されるURLであったため、同一サイトの記事については、URL毎に記事を分割するのではなく、最も短いURLをその記事を代表するURLとして利用する。

³<http://www.world-newspapers.com/>

これらの変換を行うことにより、新聞記事の本文やタイトルに関する情報以外については、NSContrast とほぼ同等のデータを作成することが可能となった。

3 分析事例

3.1 記事データベースの構築

本論文では、GDELT GKG データの 2013 年 4 月 1 日～2013 年 12 月 25 日のデータを用いて、分析実験を行った。2.3 節で述べたデータ変換を行うことにより、11,177,775 件の記事を得た。表に各地域ごとの記事数を示す。今回、提案している国名の付与のアルゴリズムは、まだ十分でないため、約 38%(4,280,168) の記事について、国名の推定を行うことができなかった。

表 1: 各地域ごとの記事数

アメリカ合衆国	2,933,282	アジア	1,295,274
ヨーロッパ	1,258,470	中東	343,595
アフリカ	392,768	オセアニア	384,462
北アメリカ	254,204	南アメリカ	35,552
分類不能	4,280,168		

3.2 分析事例

まず、NSContrast の国別注目単語比較の機能を用いることにより、それぞれの地域で注目されている単語や固有名詞間の比較を行うことが可能となる。図 1 は、2013 年 9 月 29 日における人名に対する比較で、バースト度の高い順に、名前が表示される。字の大きさは、各地域での相対的な記事数 (全体の記事のうち、固有名詞を含む記事の割合) を表している。また、背景の赤色は、その人名が特定の地域のみで注目が高いことを示している。

赤色は特定の地域のみで注目されている人名で、最も色の濃いところが、1 つの地域のみで注目されていることを示し、色が薄くなるほど、注目されている地域が増えることを示している。

また、この注目度については、1 列目は、話題の継続期間を考慮したもので、2 列目は、話題の継続期間を考慮しないものである。継続期間を考慮しない場合には、対象とする日付の各単語のバースト度によりランキングを行い、継続期間を考慮する場合には、単語が継続してバーストしている場合に、バースト開始時からの各々のバースト度を加算したものをを用いてランキングを行う。

この表から、「ronald bechtold (Mr. Ronald Bechtold:アメリカ国防総省の首席情報官)」が、世界中で

注目されていることが分かる。また、同日、「nawaz sharif(Mr. Nawaz Sharif:パキスタンの首相)」がアメリカで 10 位以内の注目されたことが分かる。

「nawaz sharif」の名前をクリックすると、「nawaz sharif」に関する注目度の情報ウィンドウが表示される(図 2)。このウィンドウには、各国での注目度の順位、対応する記事数、注目されている期間などが表示される。赤い文字で示された期間は、その期間において、継続期間を考慮した注目度が 10 位以内に入ったことがある期間を示している。

この表から、アジアでは、6 月頃に、10 位以内に入る程度に注目されていたことが分かる。

また、関連語の関係性分析を行うことにより、「nawaz sharif」を含む記事に特徴的に現れる他の人物を表示することができる(図 3)。赤いノードは、他の地域では、注目されていないが、特定の地域では、比較的注目されている人物を示している。

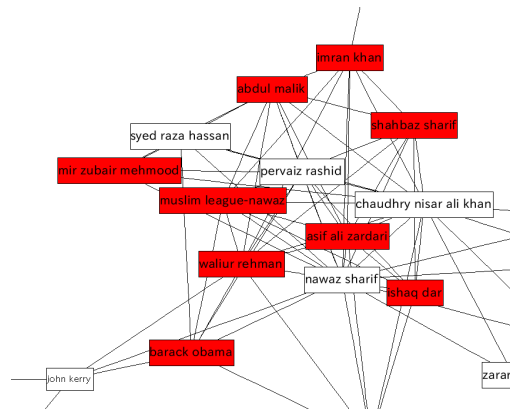


図 3: Results of the term collocation analysis

「imran khan(Mr. Imran Khan: パキスタンの政治家で、元クリケットの選手)」は、他の地域に比べて、アジアのニュース記事において、良く、「nawaz sharif」と同じ記事に登場することが分かる。この様に、特定の地域に限定した分析などを行うことで、世界的には知られていないが、特定の地域では、著名な人物などを見つけ出すことが可能となる。

最後に、この「nawaz sharif」に対して、多観点分析を行う。図 4 は、多観点分析のインターフェースで、左半分のウィンドーは、対象とする記事を観点ごとに絞り込むための検索式を作成するための領域である。この図では、対象となる人物が、「nawaz sharif」で、分析対象とする期間を「2013 年 4 月 1 日から」に設定している。

右側のグラフは、時系列グラフであり、Graph1(左上)が地域、Graph2(右上)が国名、Graph3(左下)が人名、Graph4(右下)が賛否を表している。このグラフから、「nawaz sharif」は、主に、アジア、特に、パキスタンとインドの記事に多く現れていることが確認できる。

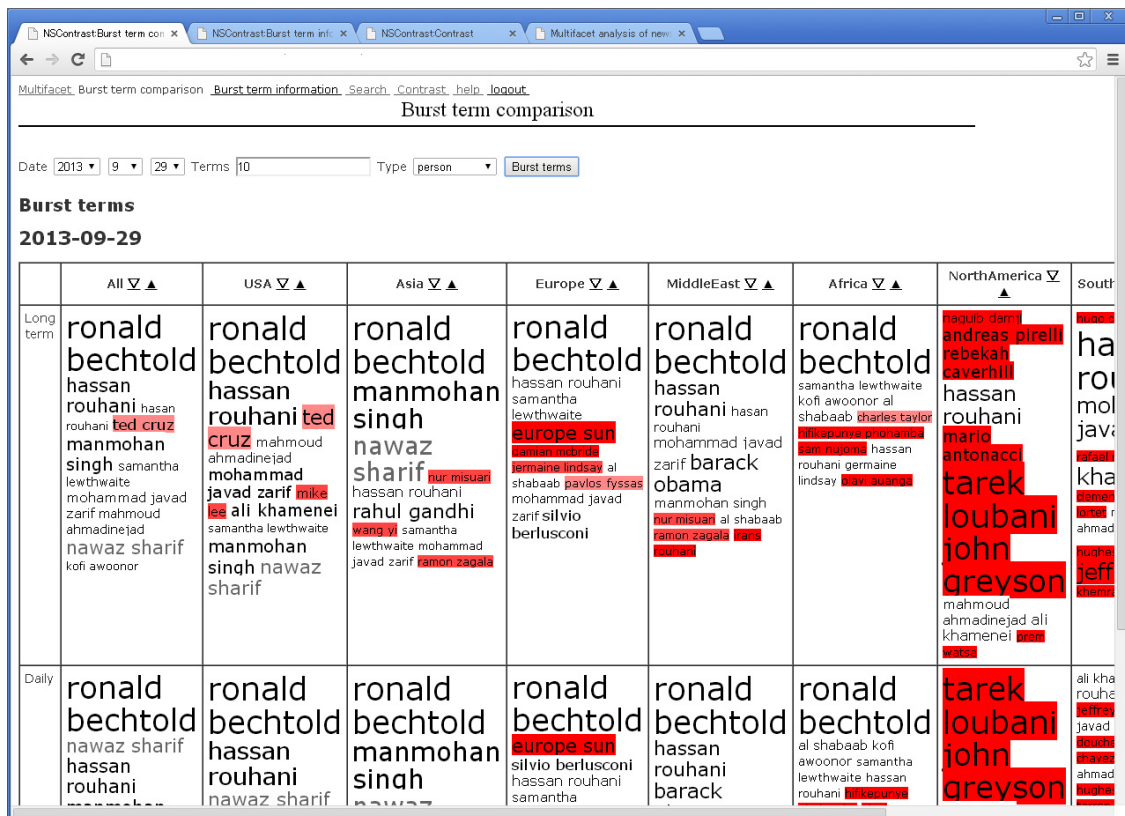


図 1: 国別注目単語比較

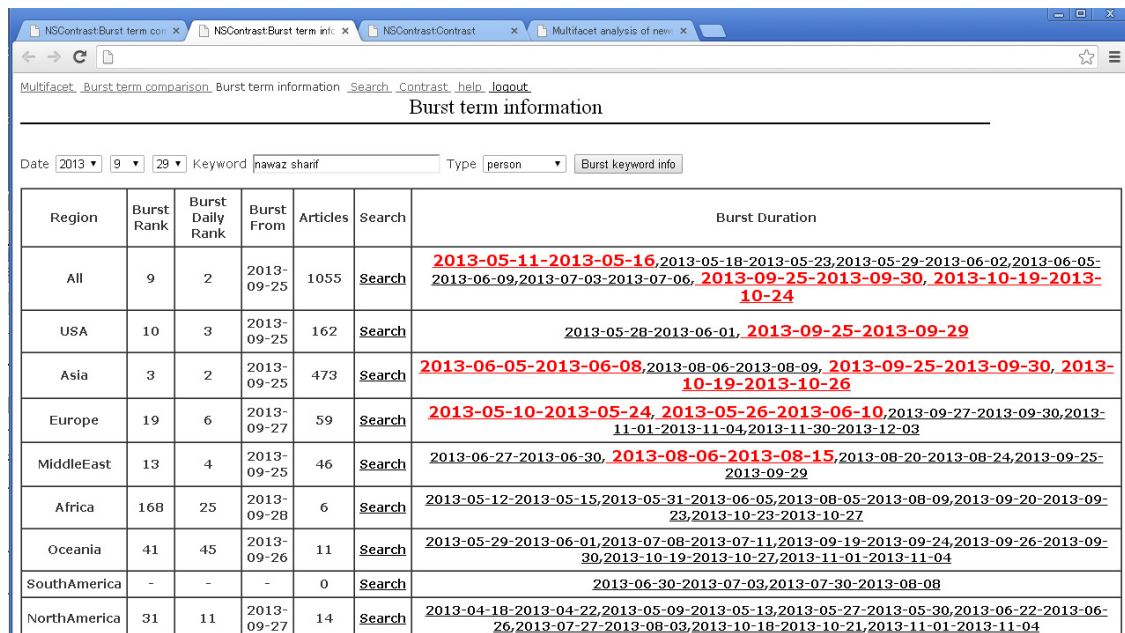


図 2: Nawaz Sharif 首相に対する注目度の情報



図 4: Nawaz Sharif を対象とした多観点分析

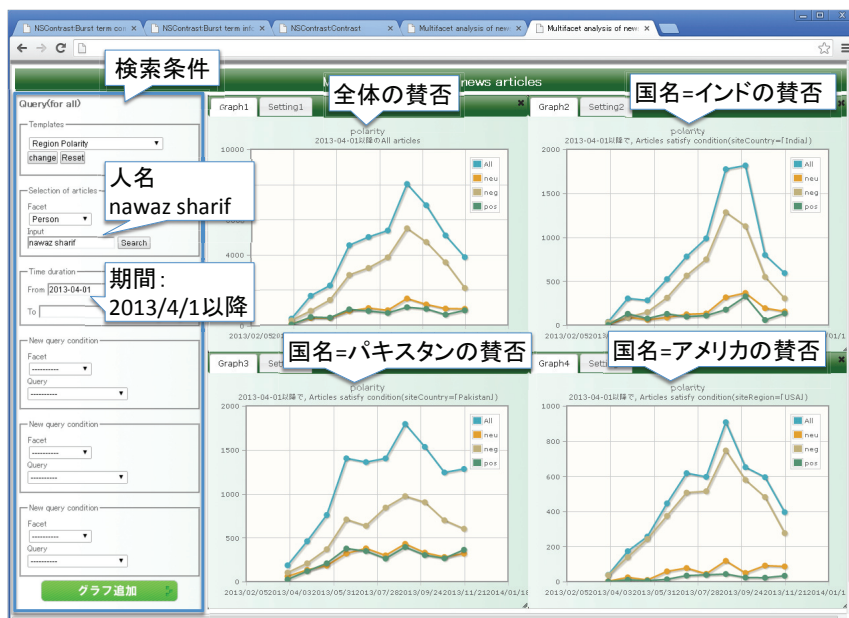


図 5: Nawaz Sharif と Imran Khan を対象とした多観点分析

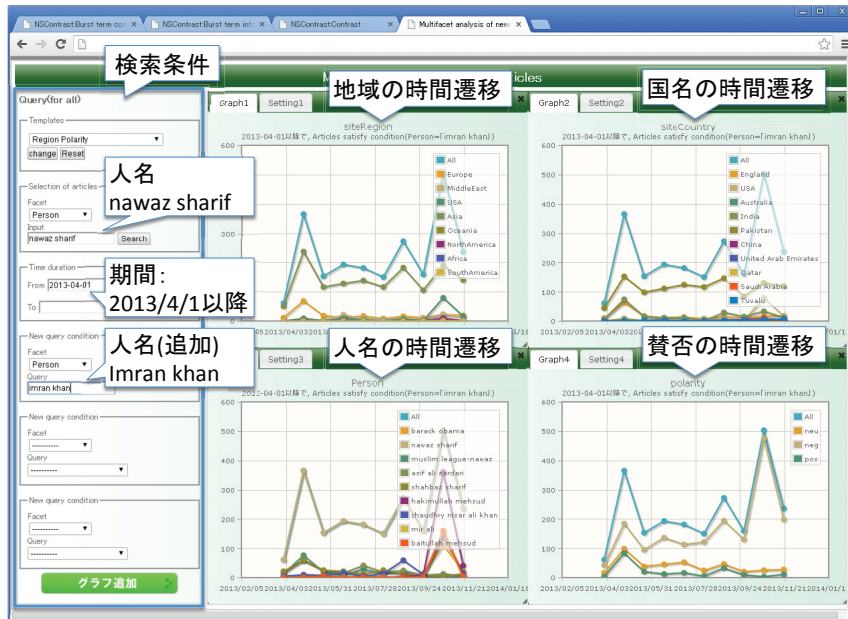


図 6: Nawaz Sharif の賛否に関する比較

この分析ウィンドウにおいて、検索式を修正すると、条件を満たした対象記事に対して、グラフが更新される。図5は、検索式に、先ほどの分析で得た人物「imran khan」を追加した場合を示す。このグラフから、6月の段階では、「imran khan」は、アジアとヨーロッパでは、それなりに注目されていたが、アメリカでは、あまり注目されていないことが分かる。しかし、11月になると、この「imran khan」は、「アメリカの無人攻撃機を利用したタリバンへの攻撃」に反対することで、アメリカでも注目されることになる。この様に、アメリカのニュースサイトを見ていただけでは、突然、注目された人物という事になるが、NSContrastのように、世界のニュースサイトを比較分析していくシステムを用いることにより、このような自体に対しても、より多角的な分析ができるようになると考えている。

また、この多観点分析システムは、様々な観点に関する情報を並べて提示するだけでなく、異なる条件の検索結果の比較を行うことも可能である。図6は、「nawaz sharif」を含む記事における賛否に対する様々な国の比較を行った結果である。Graph1(左上)が全てのニュースサイト(追加検索条件なし)、Graph2(右上)がインド(サイト国=インド)、Graph3(左下)がパキスタン(サイト国=パキスタン)、Graph4(右下)がアメリカ(サイト国=アメリカ)の賛否の時間遷移を表示したグラフとなる。この比較から、アメリカとパキスタンにおける賛否の違いなどを読み取ることが可能となる。

3.3 考察

この分析の結果、NSContrastが提供する比較分析機能は、世界中では注目されてるとはいえないが、特定の地域では注目されているような、ローカルな情報を見つけ出すのに有用であると考えている。

しかし、より精緻な分析を行うためには、以下の点において、データを充実させる必要があると考えている。

- サイトの所属する国の判定
現時点では、38%の記事について、サイトの所属する国が正しく判定されていない。これについては、より適切な情報となるように、更新することが望ましい。
- ニュース記事のタイトル
GDELTのGKGデータは、記事を単位として作られているが、記事から抽出された情報をリストとしてみるだけでは、閲覧性が高くない。少なくとも、記事のタイトルを、別のデータとして作成した上で、表示させることが出来ると、閲覧性の向上につながると考えている。

4 まとめ

本論文では、ニュース記事の多観点分析を支援するNSContrastと、ニュース記事を基盤とした大規模なイ

ベント情報である GDELT について紹介を行い、このデータを NSContrast で用いる方法について述べた。これまでの自前で作るデータベースに比較して、非常に大規模なデータが使えるという利点はあるが、こちらの目的に合わせてデータを変換する必要があったり、不足しているデータがあると言った問題点があった。

しかし、大規模なデータを使うことで、より、実際的な分析が可能になると考えられるので、このデータの利用方法について、さらなる検討を行っていきたいと考えている。

謝辞

本研究の一部は、科研費基盤研究 (B) 25280035 により行われた。ここに記して、謝意をあらわす。

参考文献

- [1] 吉岡真治, 神門典子, 関洋平. 複数国の新聞サイトを比較分析する nscontrast の実験的分析. 情報処理学会デジタルドキュメント研究会, 2011-IFAT-103, 2011. IFAT-103-2.
- [2] Masaharu Yoshioka and Noriko Kando. Multi-faceted analysis of news articles by using semantic annotated information. In *Proceedings of the fifth workshop on Exploiting semantic annotations in information retrieval*, ESAIR '12, pp. 19–20, New York, NY, USA, 2012. ACM.
- [3] Kalev Leetaru and Philip A. Schrodt. Gdelt:global data on events, location, and tone, 1979-2012. In *ISA Annual Convention 2013*, Vol. 2, p. 4, 2013.
- [4] Yohei Seki, Noriko Kando, and Masaki Aono. Multilingual opinion holder identification using author and authority viewpoints. *Information Processing & Management*, Vol. 45, No. 2, pp. 189–199, 2009.
- [5] Jon Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pp. 91–101, New York, NY, USA, 2002. ACM Press.