トピックの差異に注目した複数新聞の比較対照分析方法の提案

北海道大学大学院 情報科学研究科
吉岡 真治
e-mail:yoshioka@ist.hokudai.ac.jp

1　緒言

現在、Google News[1]に代表されるように、多くの新聞社のサイトに代表されるニュースサイトから記事を集めて関連する記事群をまとめて表示するニュースアグリゲーションサービスが提案されている。これらのサービスを使うことにより、ユーザは、幅広いニュースサイトに掲載されている記事の一覧を得ることができる。ただし、基本的に、重要な事項は、全ての新聞において同じように報道されることが想定されるため、複数の新聞記事を読んでも、類似した情報が多く存在することになる。結果として、複数の新聞を読んだとしても、読んだ文書量に比較して、あまり、幅広い観点からの情報を得られない可能性がある。

これに対し、本研究では、各ニュースサイトにおける特徴的な観点を抽出し、その結果を提示することにより、トピック中の記事の中から幅広い観点での新聞記事の分析を支援する枠組みの構築を目指している。この目標に対して、これまでに、新聞データベース間の相関性の変化に基づくトピック分析の方法を提案している [1]。

本稿では、まず、上記のトピック分析の手法について、概略を紹介し、そのトピック分析機能を中心として、現在構築中の複数ニュースサイトの比較分析システム NSContrast の持つ機能について述べる。さらに、具体的な適用事例について述べることにより、本システムについての考察を行う。

2　相関性の変化に基づくニュースサイトごとの特徴語分析

トピックに対応するような文書群を分析する方法として、文書群中に特徴的に現れる (例えば、文書群と相関性が高い) キーワードを抽出しリストアップする方法などが多く利用される。しかし、このような文書群に特徴的なキーワードのみに注目した場合には、個々のニュースサイトごとの特徴が現れるのではなく、ほとんどのニュースサイトが共通に興味を持つようなキーワードが現れ、個々のサイトごとの特徴を見つけ出すことは困難である (図 1)。

これに対し、本研究で提案するニュースサイトの分析手法では、コントラストセットマイニングの考え方に基づく相関性の変化に注目した解析 [2] を行う。具体的には、相関性の大きなキーワードに注目するのではなく、特定のニュースサイトにおけるキーワードと文書群の相関性とそれ以外のニュースサイトにおける相関性の比をとり、その比が大きいもの (そのサイトでは、それなりに注目を浴びているトピックを表すが、他のサイトではあまり述べられていないキーワード)、その比が小さいもの (そのサイトでは、他のサイトに比べて、ほとんど無視されているトピックを表すキーワード) を特徴的なキーワードとして抽出する (図 1)。

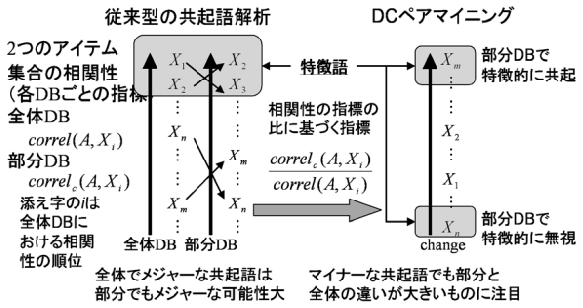

図 1: 相関性の変化に注目した特徴的キーワード分析

この考え方に基づいて、データの性質が異なることが想定される異なる国のニュースサイトの日本語版 (日本 : 3、USA : 1、韓国:2、中国:1) の約半年分の記事 (2006 年 5 月〜11 月) を全体データベースとした小規模実験を行った。この小規模実験から、例えば、「北朝鮮」というトピックに対して、従来型の共起ご会席で得られるような共通の話題を表現する語（核・ミサイルなど）を抽出するだけでなく、各ニュースサイトで特徴的に興味を持っている語（韓国の新聞では「イラク」とセットで議論されることが多い)、他の新聞に比較して興味が低い語（USA は「拉致」に興味がない）といったことが抽出可能であることを確認した。

[1]http://news.google.co.jp/

3　ニュースサイト比較分析システム

本研究では、前節で述べたキーワード分析を中心として、異なる新聞のデータを分析するニュースサイト比較分析システム NSContrast を作成している。

3.1　予備実験

本システムを構築するにあたり、次の 2 点に注目した予備実験を行った。

- 収録期間の延長による影響の有無
- 現地言語版との対応関係の分析

まず、最初に収録期間の延長による分析を行うために、新聞記事の収録期間を延長し、1 年間のデータを利用して分析を行った。データ数の増加に伴って、計算時間の増加はあったが、基本的な分析が可能であることを確認した。次に、抽出したニュースサイトごとの特徴語について分析すると、以前のものに比べ、記者の名前などの新聞社ごとの書式に伴うキーワードが増え、記事の分析に役立つキーワードの割合が減ったことが確認された。

この問題を分析するために、前回の実験で韓国の新聞において、「北朝鮮」に対する特徴語として得られた「イラク」というキーワードについて、その出現動向を調べることとした。その結果、「イラク」と「北朝鮮」の関係は、核疑惑が注目された 2006 年の後半 (前回の実験における新聞記事の収録期間) においては、かなり、注目を浴びたが、2007 年には、トピックとしての興味が薄れ、その関係について議論がされていないことが確認された。このように、経時的な変化の結果として、「イラク」というキーワードの特徴が薄れたことなどが確認された。

この問題に対応するためには、適切な期間を絞りこむ操作などを支援する必要があると考えられる。

次に、現地言語版との対応関係の分析を行うために、朝鮮日報についてのみ、韓国語版の新聞記事の収集を並行して行い、機械翻訳[2]を用いて、日本語に翻訳したデータを作成した。ただし、朝鮮日報の記事数が他の記事数に比較して非常に多く、これまでのように、一つのニュースサイトをローカルデータベースとして、残りのニュースサイト全てを全体データベースとするような方法を用いた場合に、その影響が大きくなり過ぎるという問題が発生した。

[2]高麗 2007：クロスランゲージ社

これらの問題を解消するためには、一つのニュースサイトと残りを比較する方法だけでなく、ニュースサイトのペアを用いて分析する方法などを構築する必要があると考えられる。

3.2　システムの構築

先に述べた予備実験に対する考察結果を踏まえ、次のような機能を持つシステム NSContrast を作成した。

- 情報検索システム
  サイト名や期間を限定して検索を行う。
- ニュース間の比較対照分析システム
  相関性の変化に基づく特徴語抽出を行う。従来のシステムに加え、対象期間を限定したり、比較対照するニュースサイト群の限定を可能にする。
- バースト分析
  単語の出現頻度の変化をもとに、特定の期間において注目を得たトピック語ならびに注目された期間を分析する手法であるバースト分析 [3] を行うことにより、特徴的なキーワードと期間の情報を提供する。
- トピックグラフの生成
  関連するトピックについて情報検索を行うことにより得られた文書群に対し、文書の類似度 (余弦尺度を利用) と日付情報に基づいて、類似した文書について、時間の流れに対応する形で接続したグラフを生成し、ユーザに提示する。

上記の機能の内、最初の 3 項目については、ニュースにおける特徴的な語を発見するための機能であり、最後の項目は、発見したトピックに関して、俯瞰的に情報を提示するための機能である。

3.3　検索実験と考察

本システムの有効性を検証するために、[1] と同様に収集したに二つのニュースサイトデータに対して分析実験を行った。一つ目のデータは、収録期間を変更 (2006 年 5 月～2007 年 10 月) したもの (長期) であり、もう一つは、現地言語データを取り込んだデータ (現地言語) である。こちらのデータについては、韓国語版の記事の入手期間の関係上、2007 年 8 月～10 月の 3ヶ月の短期間の記事を利用した。

表 1: 利用したニュースサイト

サイト名 (国)	URL (http:// は略)	記事数 (長期)	記事数 (現地言語)
朝日新聞 (日)	www.asahi.com/	77172	13177
日経新聞 (日)	www.nikkei.co.jp/	65915	5062
読売新聞 (日)	www.yomiuri.co.jp/	64357	6401
CNN(米)	www.cnn.co.jp/	12047	1816
朝鮮日報 (韓)	www.chosunonline.com/	21416	5062
朝鮮日報 (韓：韓国語)	www.chosun.com/	-	35919
中央日報 (韓)	japanese.joins.com/	14456	2050

利用したニュースサイト、ならびに、HTML のレイアウト解析を行うことにより、本文相当部分を抽出に成功した記事数を表 1 に示す。

3.3.1 収録期間を変更したデータに関する分析

収録期間を変更したデータに対しては、次のような手順により、分析を行った。

1. トピックとなるキーワードを決め、そのキーワードを含む記事群に対し、バースト分析を行い、特徴的なキーワードと期間を限定する。

2. 得られた期間とキーワードを用いて比較対照分析を行う。

まず、前回の実験と同様に、「北朝鮮」をトピックキーワードとして分析を行った。この分析の結果、得られた上位 5 件のキーワードとその期間は、「実験：2006.10.3〜2006.11.5」「ミサイル：2006.7.3〜2006.7.29」「発射：2006.6.30〜2006.7.29」「南北：2007.9.27〜2007.10.17」「首脳：2007.9.27〜2007.10.13」であった。これは、各々、核実験の実施の宣言日 (2006.10.3)、テポドンミサイルの発射日 (2006.7.5)、南北首脳会談 (2007.10.2〜4) に対応するキーワードである。

次に、これらのキーワードと期間を組み合わせた形で相関性の変化の分析を行ったが、その結果得られた特徴語は、収録期間を変えない場合と比べ、あまり大きな変化はなかった。また、期間のみを利用した場合でも、メジャーなトピックが与える影響が大きいため、あまり、役に立つような情報が得られなかった。

これは、バーストが起きるようなメジャーなトピックに注目しているために、さらなるトピックを表すような語を発見するのに不十分な情報しか得られなかったのが原因の一つであると考えている。

サイトごとの特徴を捉えるためには、全体のデータベースに対するバースト分析をするのではなく、個々のニュースサイトごとにバースト分析を行い、その結果を利用する方法などを検討する必要があると考えられる。

3.3.2 現地言語データを使った分析

次に、現地言語データを使った分析について述べる。。本分析においては、現地言語データを利用することにより、日本語の新聞では得ることが困難な情報を見つける可能性を示すため、やや、韓国の国内事情に関わるトピックキーワード「李明博」を用いて分析実験を行った。

表 2: 相関性の変化に基づく特徴的なキーワード

サイト名	記事数	特徴語		
朝鮮日報 (韓)	2396	操作	対策	連合
中央日報	109	現	労働党	共和党
朝鮮日報	106	祐	考える	接戦
日経新聞	7	共闘	離党	最大
読売新聞	6	和解	融和	65
朝日新聞	5	加味	座	和解

この結果、得られた特徴語を表 2 に示す。ここで、韓国語の朝鮮日報に特徴的なキーワード「操作」について調べるために、「李明博 and 操作」という検索を行った。そうすると、多くの株価操作に関連する記事が見つかった。その一部をトピックグラフで表示したものを図

これらの記事は、日本語の記事では、ほとんど記述が見つからない BBK という会社を通じた株価操作疑惑に関連する一連の報道であった。

この結果から、本システムは、現地言語を使うことにより、日本語だけでは気づきにくいトピックに関する情報を発見するのに役立つ情報を提供できる可能性があると考えている。

4 結言

本稿では、これまでに提案してきたニュースサイトの情報源の違いを分析するための比較対

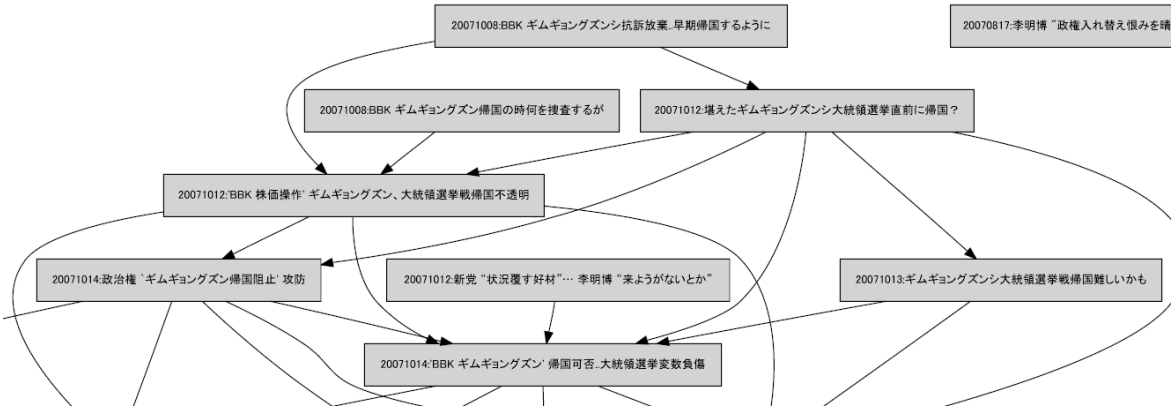

図 2: トピックグラフによる関連記事の表示

照分析の手法に基づいたニュースサイトの比較対照分析システム NSContrast について、その目的と機能について紹介を行った。本システムを用いることにより、単言語のニュースサイトを見ているだけでは気づかないような様々な観点からトピックに関する情報を分析することが可能になると考えている。

今後の課題としては、実際に具体的な分析対象を決めた上で、ユーザテストを行い、本システムが提供する情報が役に立つのかどうかについて、更なる分析を行う必要があると考えている。

謝辞

本研究を進めるにあたり、世界ニュース研究グループ (中川先生 (東大)、森先生 (横浜国立大学)、宇津呂先生 (筑波大学) ら) との有意義な議論を行った。ここに記して謝意をあらわす。

参考文献

[1] 吉岡真治. 複数のニュース源の差異を考慮したニュース分析の研究. 言語処理学会第 13 回年次大会併設ワークショップ「大規模 Web 研究基盤上での自然言語処理・情報検索研究」発表論文集, pp. 27–30, 2007.

[2] Tsuyoshi Taniguchi and Makoto Haraguchi. Discovery of hidden correlations in a local transaction database based on differences of correlations. *Data Engineering Applications of Artificial Intelligence*, Vol. 19, No. 4, pp. 419–428, 2006.

[3] 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.