NSContrast:世界ニュース比較分析システムの実験的評価

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

近年、世界中から発信される情報を用いて、世の中の動向を分析する様々なシステムが提案されている。例えば、EMM News Explorer [1]は、単純に記事を集めてきて掲載するだけではなく、その記事の発信国の情報などを含めた分析が可能となっている。

このようなシステムの一貫として、我々は、各国の新聞の興味の違いに注目して、新聞記事を分析する複数ニュースサイトの比較分析システム NSContrast を提案している [1]。このシステムでは、共起語解析の際に、トピック語と共起語の国ごとの相関性の変化に注目することにより、各々の国では、それほどメジャーではないものの、他の国との興味の違いを表すキーワードを提示する。また、本システムを実際に Web 上から収集した新聞記事のデータに適用したところ、入力したトピックを特徴づけるキーワードが抽出可能であることを確認した。

本論文では、この NSContrast の分析手法の有効性を検証するために、情報の提示手法に関する洗練化をはかると共に、複数のユーザによる利用実験を行うことにより、その有効性・問題点の分析を行ったので、それについて報告を行う。

2 NSContrast:ニュースサイト比較分析システム

本節では、本研究で提案している複数ニュースサイトの比較分析システム NSContrast について、その分析手法の基礎となる相関性の変化に基づく特徴語分析の手法と、システムの概要について述べる。

2.1 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)。

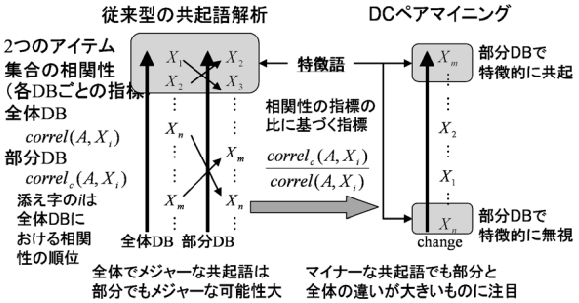

図 1: 相関性の変化に注目した特徴的キーワード分析

2.2 システムの概要

NSContrast では、日本語の新聞サイトから獲得した記事ならびに、中国・韓国の現地語新聞を機械翻訳した記事から構成される新聞記事データベースに対して、下記の機能により記事の分析を行う。

- 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。システムは、各国ごとに、相関性の高い特徴語、相関性の変化の高い特徴語・相関性の変化の低い特徴語をリストとして表示する。

[1] http://press.jrc.it/NewsExplorer/

- 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。

3 NSContrast の改良とユーザ実験

3.1 NSContrast の改良

本システムの有効性を検討するために、実際にユーザに利用実験に参加してもらい、その評価を行うこととした。この利用実験にあたって、システムに関する意見を求めたところ、以下の 3 点の問題点が指摘された。

1. 特徴語を単にリストと表示する機能だけでは、理解が難しい。
2. 特定のトピックに関連するバースト分析では、国ごとに興味のあるトピックの違いなどが理解できない。
3. 最新の記事がないと、実感を持った評価が困難である。

問題点 1 を解消するために、特徴語間の関係を考慮した表示方法の検討を行った。この特徴語の関係を示す方法では、次の 2 つの観点の情報をまとめて表示することとした。

- 特徴語間の共起関係の可視化
  特徴語間で共起関係の強いものを関連づけたグラフとして表示することにより、関連トピックを理解可能とする。
- 国ごとの興味の違いの可視化
  特徴語がどの国の記事に多く現れているかを考慮して、上記の共起関係のグラフにあわせて表示することにより、国ごとの興味の違いを可視化する。

問題点 2 に関しては、各国ごとに、バースト状態にある単語を調べると共に、バーストしている期間やバーストの度合い (どれくらい平常状態よりも語の出現頻度が高いか) を考慮したランキングを行って比較する機能を実装した。

また、問題点 3 を解消するために、表 1 のサイトから定常的に新聞記事を収集し、毎晩、定常的にデータを更新することにより、朝になると前日までの記事を利用した解析が可能なように設定を行った。表 1 の記事数は、2008 年 1 月 1 日からの記事数の総計と 1 日あたりの平均記事数 (2009 年 1 月 12 日現在) を示す。

表 1: 利用したニュースサイトと記事数

サイト名 (国) URL (http:// は略)	記事数 (総計) (1 日平均)
朝日新聞 (日) www.asahi.com/	43395 115
日経新聞 (日) www.nikkei.co.jp/	54662 145
読売新聞 (日) www.yomiuri.co.jp/	36293 96
CNN(米) www.cnn.co.jp/	7492 20
朝鮮日報 (韓) japanese.chosun.com/	19287 51
中央日報 (韓) japanese.joins.com/	14705 39
人民網 (中) j.peopledaily.com.cn/	14306 38

3.2 改良したシステムによる解析事例

本システムの機能を具体的な解析事例と共に説明する。

1. バースト解析による報道量の違いの分析

ユーザは、まず、興味のあるトピック語をシステムに入力する。システムは、トピック語を含む新聞記事を検索すると共に、記事の出現頻度をベースとしたバースト解析の結果を表示する。原油をトピックキーワードとし、6 月 23 日時点での解析結果を図 2 に示す。

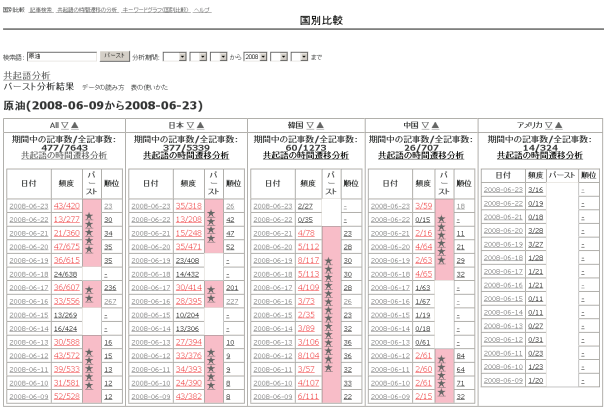

図 2: バースト解析結果の国別比較

この表では、日本のニュースサイト、韓国のニュースサイト、中国のニュースサイト、アメリカのニュースサイトを各々ひとまとめとし、各々の期間での検索語を含む記事数を表示している。また、日付に対応した欄が赤く (薄く) 塗り潰されており、バーストという欄に★がついている期間が検索語がバーストしている期間である。この表から各国における注目度の違いを理解することができる。

2. 共起語解析による国ごとの特徴の分析

次に、これらのバースト情報に基づいて、共起語の対照比較を行う。昨年度のシステムで用いたような表形式の表現では、特徴語として抽出された語の間の関係が不明確であったため、本システムでは、特徴語間の共起関係の強さ、各国の新聞と特徴語の共起関係の強さをもとにリンクを設定したバネモデルによる共起語の可視化を行った。図 3 に先ほどの原油のトピックにおいて、全体的にバーストしている 6/19〜23 日の記事を利用した分析結果を示す。

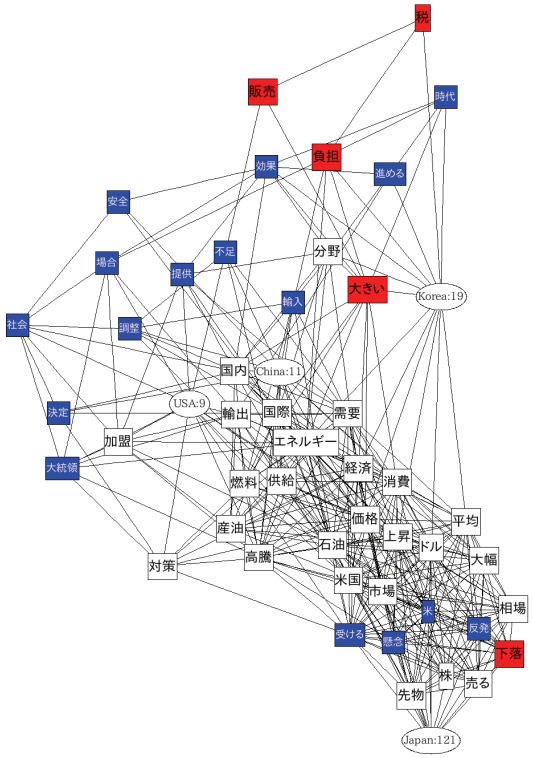

図 3: 特徴的な共起語間の関係グラフ

このグラフでは、相関性の変化が大きい共起語がどのようなメジャーなトピックに関係しているかを可視化するために、各国ごとに、相関性の変化が大きい共起語、相関性の高い共起語を各々10 個ずつ選んで表示している (各国語との重なりがあるので、全体では、80 個より少ない)。白で表示されたものは少なくとも一つ以上の国で、相関性の高い語でメジャーなトピックを表す語である。濃い背景に白字で表された語は、相関性の変化が高い語で、特定の国の近くに配置される。濃い背景に黒字で示された語は、相関性も高く、相関性の変化も高い語で、その国においてメジャーでかつ、他の国ではメジャーでない特徴語を示している。図 3 からは、日本では、相場が下落し始めたことを、いち早く報道しており、韓国では、税負担の可能性が議論されているといった違いを見ることができる。

3.3 ユーザ実験 1：簡単なアンケート分析

本システムの有効性を検証するために、科学研究費特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」における様々なシステムの実験的評価の一環として、情報系の学生を中心とした利用実験を行った。この実験は、簡単なシステム操作結果に基づいて、表 2 に示すアンケートの質問による評価を行うと共に、自由回答のコメントを回収した。

表 2: システムの評価結果

質問	Yes	No
日本と他国の報道の違いを気づくのに役立つ情報が提示されましたか？	34 (57%)	26
検索項目に関連する記事内容を理解するのに役立つキーワードが提示されましたか？	37 (61%)	24

その結果、約 60%の利用者から、本システムにより、日本と他の国の報道の違いを理解するのに役立つ情報が得られたという回答を得た。一方で、日本の記事数に比べて中国や韓国の記事数が少ないため、中国や韓国の細かな国内事情を把握するまでの能力がないという指摘があった。

3.4 ユーザ実験 2：レポート作成課題の実践

ユーザ実験 1 では、システムの持つ機能を大まかに分析してもらうことが目的出会ったのに対し、ユーザ実験 2 では、NSContrast の作成意図を考慮した利用シナリオを想定した実験を行った。具体的には、特定の事象に関する各国の報道の違いをレポートとしてまとめるという作業のために、本システムを利用するというシナリオを想定した実験を行った。

本実験を行うにあたり、ユーザ実験 1 の問題に対処するために韓国語、中国語の新聞については、中国・韓国の現地語の新聞に対し、機械翻訳 (クロスランゲージ社の Webtranser) を適用することにより、NSContrast のデータベースの拡充を行った。追加したニュースサイトの情報を表 3 に示す。

表 3: 追加したニュースサイトと記事数

サイト名 (国) URL (http:// は略)	記事数 (総計) (1 日平均)
朝鮮日報 (韓) www.chosun.com/	54154 144
新華社 (中) www.xinhuanet.com/	319419 847

今回の実験の目的は、以下の三点についての情報を得ることにある。

1. NSContrast の機能のうち、どの機能が実際の分析において有用なのか。
2. 機械翻訳システムが与える影響について。
3. 多言語の新聞記事データを利用することで、得られる国ごとの特徴のデータとしては、どのようなものがあるのか。

この目的のため、各被験者には、設定した課題に対する最終のレポートを提出するだけではなく、作業記録 (システムのどのような機能を使って、どのような情報を発見したのか) と、システムの評価を同時に提出してもらうこととした。

また、機械翻訳システムが与える影響を分析するためには、元々の言語の記事との対応関係を理解できることが望ましいので、今回の実験では、日本語に加え、中国語もしくは韓国語の少なくともどちらかは理解できる 4 名の方を被験者とした。

各被験者には、日中韓に関する興味が違うであろうと考えられるトピックを各々に設定してもらい、実験を行った。各被験者が設定した課題は下記の通りである。

被験者 1 北朝鮮、金融危機、チベット、メラミン、秋葉原

被験者 2 四川地震、インドテロ

被験者 3 北朝鮮、金融サミット、オリンピック、李明博大統領、サムスン

被験者 4 狂牛病、竹島、ニート

各々の被験者の作成したレポートから、各国の新聞の報道の違いを見つけるためのキーワードが本システムの特徴語分析から発見できたことが報告されると共に、下記に記す問題点が指摘された。

- 表記の違いの影響を強く受ける。
  特定のニュースサイトの記事において、他の新聞と異なる特有な表現が含まれる場合や、翻訳システムが一貫して間違いを行った場合に、その語が特徴語として抽出されやすい。ただし、この特徴語を利用することにより、記事の再検索のためのキーワードが得られる場合がある。
- 共起語間の関係の可視化について
  共起語の可視化自体は有効であるが、国との関係については、必ずしも、適切な配置であるとはいえない。

特に、前者の問題は 4 名の被験者が共通して指摘していた問題であり、今後のユーザ実験を行う前に解決すべき問題であると考えられる。

一方で、バースト解析の結果の比較は、新聞報道の遅れ (例えば、メラミンによる食品汚染の問題における、中国の報道の遅れ) などを理解するのに有効であった、といったシステムの利用方法における工夫についても報告された。

これらについても、さらに考察を行うことにより、システム改良の方針を固めたいと考えている。

4 結言

本稿では、これまでに提案してきた、新聞記事を分析する複数ニュースサイトの比較分析システム NSContrast のの目的と機能について紹介を行うと共に、異なるタイプのユーザ実験を行うことにより、システムの有用性を検討した。

本システムでは、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情報を提示する可能性を示している一方で、新聞ごとの表記の違いなどに対する影響を強く受けることが確認された。この問題は、機械翻訳のデータを使った場合に顕著に現れる。今後は、Wikipedia のエントリーの情報を利用した同意語の正規化などを行うことにより、表記の違いの影響の軽減を試みる予定である。

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)、宇津呂先生 (筑波大学)、福原先生 (東大) ら) との有意義な議論を行った。また、科学研究費特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」の関係者には、ユーザ実験に協力をしていただいた。ここに記して謝意をあらわす。

参考文献

[1] 吉岡真治. 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. 言語処理学会第 14 回年次大会発表論文集, pp. 592–595, 2008.

[2] T. Taniguchi and M. 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] J. 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.