Comparative Analysis of GDELT Data Using the News Site Contrast System

Masaharu Yoshioka
Hokkaido University
N14 W9, Kita-ku, Sapporo-shi,
Hokkaido, 060-0814, Japan
yoshioka@ist.hokudai.ac.jp

Noriko Kando
National Institute of Informatics
2-1-2, Hitotsubashi, Chiyoda-ku,
Tokyo, 101-8430, Japan
kando@nii.ac.jp

Abstract

The News Site Contrast (NSContrast) system analyzes news articles retrieved from multiple news sites based on the concept of contrast set mining. It can extract terms that characterize different topics of interest across news sites, countries, and regions. In this study, we used NSContrast 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 Introduction

It has become possible to access a wide variety of news sites from across the world via the Internet. Each news site has its own culture and interpretation of events, so we can obtain a greater diversity of information than ever before by using multiple news sites. Opinions and interests expressed in news articles vary across countries, and we can obtain different points of view regarding a topic if we access news sites from different countries. For example, Asian, European, and American news sites share some common views on diplomatic issues related to North Korea, as well as having their own characteristic opinions. Therefore, it is important to clarify the characteristics of each specific news site when analyzing events reported by multiple sites.

The News Site Contrast (NSContrast) system was developed to analyze the characteristics of news sites [YK12]. However, since it is not easy to construct news databases from different countries, NSContrast only uses small numbers of news sites from East Asian countries (Japan, China, Korea) and the USA to characterize the differences between them.

Recently, a Global Database of Events, Language, and Tone (GDELT) [LS13] [1] was released. This database is based on larger numbers of news sites from all over the world and it contains extracted metadata information from news articles. In this paper, we propose a method to utilize GDELT to analyze the characteristics of news article from different countries and regions by adding country and region information for the news sites in the database. By using these data, we can compare news articles from various countries and regions (e.g., USA, Asia, South America, and Africa) worldwide instead of our original small database of news articles. We also present examples of analyses performed using the NSContrast system.

2 NSContrast

2.1 System description

NSContrast employs the following four methods to analyze news articles.

- **Burst analysis** [Kle02] identifies the daily burst terms and the regional distribution of a specific bursty term. (Figure 1)
- A **term collocation analysis graph** shows relationships among collocated terms and the given query. NSContrast uses highly collocated terms from all regions based on contrast set mining and ordinal collocation analysis. These collocation terms are visualized with a spring model using fdp in Graphviz.[2].
- A **news article retrieval system** is used to understand the meanings of the terms in the collocation analysis and the burst analysis.
- A **multifaceted interface for analyzing news articles**.

The system uses multiple facets (e.g., keyword,

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos, and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop at ECIR, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

[1] http://www.gdeltproject.org/

[2] http://www.graphviz.org/

named entity, polarity, news site, and country) to analyze news articles. The interface supports the construction of structured queries that use one or more facets, where the facet information can be represented using various styles (e.g., time sequence graph, table, or bar chart). (Figure 2)

2.2 Data conversion

To apply NSContrast to the analysis of GDELT data, it was necessary to convert the GDELT data into news article data. There are two databases in GDELT: GDELT Event and GDELT Global Knowledge Graph (GKG). GDELT GKG is a database based on a raw output format of the original news articles for constructing the GDELT Event database. Because the GDELT Event database does not have detailed original news article sources, GDELT GKG was used for NSContrast.

GDELT GKG was constructed by extracting the following metadata information from the original news articles: DATE, THEMES, LOCATIONS, PERSONS, ORGANIZATIONS, TONE (as a real value; 0 means neutral), CAMEOEVENTIDS (references to the GDELT Event database), SOURCES, and SOURCEURLS. When there are two or more articles that share all name sets (THEMES, LOCATIONS, PERSONS, and ORGANIZATIONS), those news articles are aggregated as one datum and SOURCES and SOURCEURLS have multiple entries. Example of SOURCES and SOURCEURLS information for one datum in January 19, 2016 are shown below.

SOURCES punchng.com; punchng.com; onlinenigeria.com; onlinenigeria.com

SOURCEURLS http://www.punchng.com/25909-2/, http://www.punchng.com/i-am-resolved-to-better-lagos-ambode/, http://news2.onlinenigeria.com/news/general/453949-i-am-resolved-to-better-lagos-%E2%80%93w-ambode.html, http://news2.onlinenigeria.com/news/general/453949-i-am-resolved-to-better-lagos-ambode.html

Two types of multiple SOURCEURLS are shown above. In one, almost the same content has a different URL for the same news site (the first two URLs and the last two URLs above) and the other is a different URL with different news sites (the first and third URLs).

Most of the former cases are simply URL variations of the same content; e.g., the first URL is redirected to the second URL and the third URL is a variation of the fourth URL (the URL encoding of "%E2%80%93w" is "-" for UTF-8). It is better to select one of them for deduplication. The latter cases are meaningful for representing the importance of the contents, because different news sites have selected the same content for their sites.

By using these metadata, the following information was constructed for NSContrast.

Date Date of the article.

Person, Organization, Location Lists of people, organizations, and locations extracted from the article using the GDELT GKG.

Polarity We classified articles into three types (positive, negative, and neutral) to simplify the analysis of the polarity information. The tone extracted by the GDELT GKG was used for classification (tone > 1: positive; tone < -1: negative; other: neutral).

Site Site information extracted from GDELT GKG. To count the number of articles from different news sites, we duplicate one datum for each site. However, if there are two or more entries for the same news site information, one of these entries is used for deduplication. In the above example, one datum is duplicated for "punchng.com" and "onlinenigeria.com."

URL URL for the original news article. When there is one URL for a site, the corresponding URL is used for each site. However, when there are two or more URLs for a given news site, the shortest URL is selected for each news site (e.g., http://www.punchng.com/25909-2/ for punchng.com).

SiteCountry We constructed a database of news sites to identify their countries of origin. We used `http://www.world-newspapers.com/` to extract these relationships. For "BBC monitoring," we used "United Kingdom" as the site country for the news site. In addition, if news sites used country code top-level domains (e.g., .jp for Japan), we used this domain information to estimate the site country. Finally we used a geolocation service [3] to estimate the site country by using the IP address of the top domain. However, the country was left blank if we could not obtain appropriate location information from the geolocation service.

SiteRegion Countries were grouped into the following eight regions: USA, Asia, Europe, Middle East, Africa, Oceania, North America (excluding USA), and South America. News articles that lacked site country information were categorized as Unclassified.

We could use all of these information types other than the URL to perform multifaceted analyses.

3 NSContrast with GDELT

We set up our system based on the GDELT GKG from July 20, 2015 to January 19, 2016. Using the data conversion process described above, we extracted 31,584,327 articles from 70,781 news sites.

[3] https://freegeoip.net/, http://ip-api.com/, and http://ipinfo.io

First, we present information related to the country and region estimation. Because our manually constructed news site list is small, only 2201 sites (8,555,263 articles) were identified by using this information. Table 1 shows the number of articles (sites) by the top-level domain of URLs (Top 6). Because 81.2% (47,259/70,781) of news sites and (71.9% (22,716,591/31,584,327) of articles have .com as their top-level domain, only 10,139 sites (5,671,259 articles) were identified by their top-level domain.

Table 1: Number of articles (sites) for top-level domains (Top 6)

.com	22,716,591 (47,259)	.au	2,623,813 (1048)
.uk	1,682,960 (2705)	.org	1,029,232 (8049)
.net	645,996 (3015)	.ca	326,184 (1008)

Finally, by using the geolocation service 57,459 sites (16,816,980 articles) were identified. As a result, most of the sites (98.6%: 69,799/70,781) and articles (98.3%: 31,043,442/31,584,327) were classified into countries and regions.

Table 2 shows the number of articles for each region. From this table, news articles from the USA were dominant in the database (61.6%: 19,443,005/31,581,063). In contrast, there were only 903,811 articles from North America excluding the USA. With such unbalanced numbers of articles, making a category North America including the USA is almost equivalent to USA alone. Therefore, we divided North America into the USA and North America (excluding USA).

Table 2: Number of articles for each region

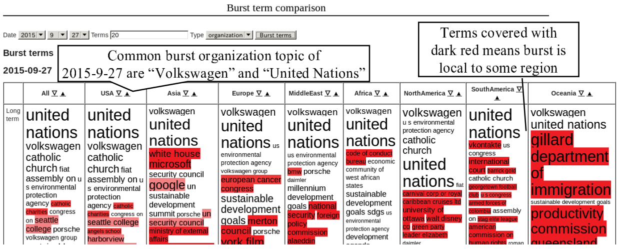
USA	19,443,005	Europe	3,696,359
Oceania	2,962,792	Asia	2,891,865
North America (excluding USA)			903,811
Africa	726,626	Middle East	373,411
South America	45,573	Unclassified	537,621

Our multifaceted analysis interface was used to compare the results with different query conditions. Figure 2 shows a time-sequence graph of polarity in different countries: all countries (upper left), China (upper right), the USA (lower left), and Europe (lower right). These graphs were constructed by adding new query conditions when selecting the data. For example, the graph for China uses news articles that included "Asian Infrastructure Investment Bank" (AIIB) as the organization, an article date ≥ July 20, 2015, and the SiteCountry = "China."

This figure shows that there were many positive articles about AIIB in China. Europe was slightly positive than the USA. This information reflects the attitudes to AIIB in these countries (or regions).

4 Conclusion

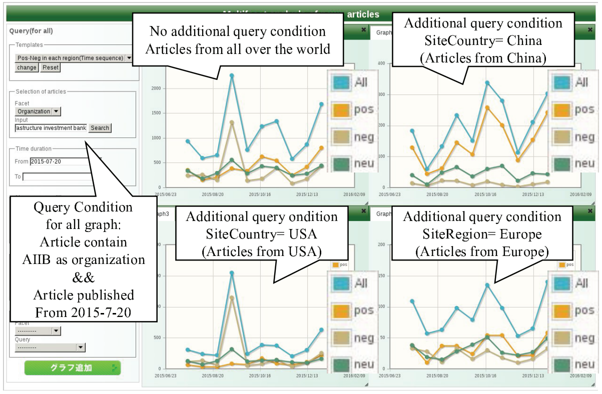
In this study, we have analyzed the characteristics of GDELT data and propose a data conversion process to utilize this information for NSContrast. In this conversion process, we conducted deduplication of news article URLs and added source country and region information to analyze the characteristic differences between them. Because of the large coverage of news sites, the system can conduct comparative analyses of various countries and regions by using large numbers of news articles from different news sites. However, for future work, it may be better to check the appropriateness of the estimated country by using a geolocation service.


Burst term comparison

Common burst organization topic of 2015-9-27 are "Volkswagen" and "United Nations"

Terms covered with dark red means burst is local to some region

Figure 1: Burst analysis results on 2015-9-27


No additional query condition
Articles from all over the world

Additional query condition
SiteCountry= China
(Articles from China)

Query Condition
for all graph:
Article contain
AIIB as organization
&&
Article published
From 2015-7-20

Additional query ondition
SiteCountry= USA
(Articles from USA)

Additional query condition
SiteRegion= Europe
(Articles from Europe)

Figure 2: Multifaceted analysis for the query "AIIB"

Acknowledgement

This work was partially supported by JSPS KAKENHI Grant Number 25280035.

References

[Kle02] Jon Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pages 91–101, New York, NY, USA, 2002. ACM Press.

[LS13] Kalev Leetaru and Philip A. Schrodt. Gdelt:global data on events, location, and tone, 1979-2012. In *ISA Annual Convention 2013*, volume 2, page 4, 2013.

[YK12] Masaharu Yoshioka and Noriko Kando. Multifaceted analysis of news articles by using semantic annotated information. In *Proceedings of the fifth workshop on Exploiting semantic annotations in information retrieval*, ESAIR '12, pages 19–20, New York, NY, USA, 2012. ACM.