A Framework for Blog Data Collection: Challenges and Opportunities

Muhammad Nihal Hussain, Adewlae Obadimu, Kiran Kumar Bandeli, Mohammad Nooman, Samer Al-khateeb, Nitin Agarwal*

*Maulden-Entergy Endowed Chair and Distinguished Professor of Information Science
Collaboratorium for Social Media and Online Behavioral Studies (COSMOS)

nxagarwal@ualr.edu

University of Arkansas at Little Rock
Agenda

• Literature Review
• Data Collection Methodology
• Data Description
• Blogtrackers Analysis
  – Posting Frequency
  – Sentiment Analysis
  – Influence
• Conclusion
• Future Work
**Blog Crawling:** These crawling techniques scrape only a fraction of the posts from the blogs

- Berger et al. (2011) crawled blogs using RSS feeds and DOM parser.
- Woo et al. (2016) extended the work of Ginsberg et al. (2009) and used blogs to estimate the outbreak of influenza in South Korea.

**Deep Web crawling:** These efforts focus primarily on deep web.

- Zhao et al. (2016) designed a two-stage SmartCrawler framework for scraping the deep web by reverse searching the known deep web sites and ranking the collected sites.
- Wang et al. (2017) developed an crawler that exhaustively scraped textual data from ranked deep web sources with minimal cost.
Blog tracking tools: Due to the complexity and inefficiency of blog data collection many blog tracking tools have shutdown.

- BlogPulse – developed by IntelliSeek. provided search and analytical capabilities, automated web discovery for blogs, show the trends of information, and monitor the daily activity. It was shutdown in 2012.

- Blogdex – was developed at “The Media Lab” at MIT. It provided a time-weighted ranking of trending posts. This feature could be used to identify hot-topics in blogosphere. It was shutdown in 2006.


- Technorati – used proprietary search and ranking algorithm to provide blog index and search engine. It did not provide blog monitoring or analytical capabilities. Discontinued blog index in 2014.
Data Collection Methodology

Step 1: Blogs Exploration

Step 2: Training the Crawler

Run the Crawler to Extract the Data

Standardization/Data Processing/Metadata Extraction

Data Validation and Verification

Removal of Missing Values

Step 3: Cleaning & Storing the Data

Blogtrackers Database
Blog Exploration

- Structure of Site
  - Page navigation
  - Archives
- Single-author/Multi-author
- Attributes crawled
  - title
  - author
  - date
  - content
  - comments
  - tags

NATO Summit in Brussels: Armament and war but also „we will not be silent“

At the NATO Summit in Brussels US President Trump wants to collect imaginary debts and the European heads of state reiterated their willingness to spend 2% of GDP on armament. A large wave of armament is coming, for Germany 2% military spending means an increase from 37 billion Euro to 69 billion Euro, for Europe an increase of 200 billion to more than 300 billion Euro. The Europe of crisis and joblessness should pay for preposterous interventions and wars while the military industry rejoices. And there will be cuts in education, science, health care and environment.

The Summit agreed on NATO’s participation in the so called “war on IS”. In reality this “war on IS” is a series of illegal wars in which Germany now will participate more actively than before. This so called “war on IS” brings terror to a whole region and therefore is provoking terrorist acts. An even more brutal war of bombs will follow with NATO’s decision. The majority of the victims are the innocent. Each innocent death strengthens IS and allied “groups of terrorists”. The war in and for Syria will be increased turning the whole region into an even more difficult to control tinderbox. The continuous war on terror will also fail because wars do not solve problems but solely increase them by destabilizing whole societies, countries and regions.

The “3 Cs” of the summit – cash, capabilities, contributions – can be easily translated into more money for modern weapons and for more wars worldwide.

In Brussels was more than a city of the active and cool warrens.

More than 12,000 protesters were marching through Brussels in a colorful, young, broad, impressive and loud demonstration on 24 May. For many hours a trail of peace was moving through the city of Brussels. The actions were creative, musical and international. Many countries with many slogans – an impressive atmosphere. Everywhere “no to NATO” was echoing through the streets. The common demeanor for peace activists, critics of globalization, gender activists, and environmentalists was a clear “no” to further armament. This demand forged a broad coalition of resistance with diverse political backgrounds.

More than 200 people attended to counter summit of the Belgian and international peace movement on 25 May. It was characterized by its internationalism and by discussions on challenges to peace. In an atmosphere of solidarity and mutual tolerance several common points became obvious:

- The peace movement’s challenge is to work against the new round of armament towards the 2% of GDP and to struggle for real disarmament: disarmament for development and disarmament for the solution of the social and global challenges. The participants of the conference were prepared to work more intensively towards achieving these challenges.
- The UN ban treaty of nuclear weapons must become reality. There needs to be nuclear disarmament instead of the modernization of nuclear weapons. This is the message to all nuclear weapons states. And Europe finally has to become free of nuclear weapons.
- Cooperation instead of confrontation is not only meaningful but necessary and possible. Especially with Russia. All enemy constructions and bashings serves the preparations of war.
- An end to wars of intervention – in Mali, Afghanistan and many other places of NATO’s wars – is the condition and requirement for a peaceful and just development of the world.
To train the crawler:

- Start with seed URLs
- Page navigation techniques
  - Navigation to next page/next set of posts
  - Navigation to actual/each post
- Extraction pattern
  - DOM pattern for each attribute
- Meta data
  - Permalink
  - Data collection date
Cleaning and Storing

Cleaning using JAVA
• remove noise from the content
• Data standardization and manipulation

Data extraction
• URLs and Domains using JAVA
• Entity extraction, their types and directed sentiment towards them using AlchemyAPI
• Language of posts using AlchemyAPI
• I.P. address, hosting location, analytics or tracking ID and any related sites using Maltego
• Sentiment and tonality using LIWC 2015
Challenges faced during data crawling

- **Changing blog structure** – Blog site owners can change their blog structure. The crawler needs to re-trained for the new structure of the blog site.

- **Noise** – Irrespective of how well a crawler is trained, noise is always crawled. Social media plugins like Facebook share plugins, Twitter share plugins, etc. and advertisements from the blog site could be crawled as JavaScript or HTML code.

- **Limitations of WCE** – WCE sometimes fails to crawl dynamic pages that are loaded using JavaScript.

- Some of the technologies used during data collection and extraction steps are not free and require commercial licensing.
**Database Statistics**

- 257 blogs
- 213,785 blog posts
- More than 1,461,062 links
- 4,058,340 entities
- Earliest post from February 1993
- Last crawled post from May 2017

<table>
<thead>
<tr>
<th>Language</th>
<th>Blogs</th>
<th>Blog Posts</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>191</td>
<td>164082</td>
</tr>
<tr>
<td>Spanish</td>
<td>41</td>
<td>18474</td>
</tr>
<tr>
<td>Russian</td>
<td>25</td>
<td>5438</td>
</tr>
<tr>
<td>French</td>
<td>25</td>
<td>235</td>
</tr>
<tr>
<td>German</td>
<td>22</td>
<td>1411</td>
</tr>
<tr>
<td>Italian</td>
<td>16</td>
<td>579</td>
</tr>
<tr>
<td>Polish</td>
<td>23</td>
<td>5368</td>
</tr>
<tr>
<td>Danish</td>
<td>11</td>
<td>23</td>
</tr>
<tr>
<td>Latvian</td>
<td>11</td>
<td>3794</td>
</tr>
<tr>
<td>Estonian</td>
<td>7</td>
<td>810</td>
</tr>
</tbody>
</table>

Top 10 languages
Blogtrackers
(http://blogtrackers.host.ualr.edu/)
Posting Frequency for Anti-NATO blogs from January 2015 to December 2016

These peaks correspond to intensive blog discussions on NATO’s Trident Juncture exercise in late 2015.
There is a growing negative sentiment within the blog discourse for NATO’s Trident Juncture exercise. The negative sentiment was indicative of the growing anti-NATO narrative in the blogs.
Influence

Influence trends for top 5 bloggers

No Bordersard had the highest influence in August 2015
Influential Posts

Blog posts of No Bordersard (most influential blogger) ranked in decreasing order of influence
Conclusion

• We introduced a semi-Automated crawler, where human trains a crawler tailored for a blog site and lets the crawler run until all the posts from the blog are collected.

• Blogosphere is a relevant information environment to study events, related discourse, and leading narratives.
Future work

• Automated crawler that identify the attributes from the HTML.

• Add content analysis features to Blogtrackers like:
  • topic modeling,
  • network analysis, and
  • cyber forensics features,

• To not only study blogs individually, but also to understand their coordination structure and information dissemination structure.
Acknowledgments

This research is funded in part by the

– U.S. National Science Foundation,
– U.S. Office of Naval Research,
– U.S. Air Force Research Lab,
– U.S. Army Research Office,
– U.S. Defense Advanced Research Projects Agency, and
– Jerry L. Maulden/ Entergy Fund at the UA-Little Rock.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.
Thank you!

nxagarwal@ualr.edu

http://blogtrackers.host.ualr.edu/