Visualizing Crisis News Briefs

Andrea Krukowski (ajk583)
Meredith McCarron (mkm318)
Ady Sevy (as8202)

Introduction

The Operations Center (OPSCEN) at UNICEF sends out via email extensive daily summary reports of recent security and crisis-related news to relevant staff members in every country in the world as well as headquarters policy staff and any other parties who may be interested (such as workers at other humanitarian or crisis-response organizations). The reports contain one-paragraph summaries of news articles, grouped by world region and country. Reports span about 10 single-spaced pages. Such large amounts of text data is very difficult for users to absorb, and comparing stories and themes over time is particularly burdensome. Headquarters staff primarily uses the briefs to browse, as reference for policy papers, and to keep track of risks to the security of the 12,000 UNICEF staff members who work in the field. Therefore, one major interest is following long-term and short-term trends on a country, region and worldwide scale. However, field staff may represent a very different use-case -- less interested in browsing or events affecting in other regions. They are more likely to focus on the events in the country where they are stationed or the events in countries nearby that may spillover into their area.

These varied use-cases for the data is one of the biggest challenges of our project -- therefore, we plan to take care to design a visualization system that meets *most* of the needs of each group. So far, we have access to some end-users (the headquarters staff who compile the reports and keep track of security risks and the headquarters staff who use the reports for policy proposals). Talking to all of the high-level field staff who receive these reports would be nearly impossible, though we are interested in being connected to one or two of these staff to gather anecdotal input.

At present, the only way to use these briefs to identify trends is for users to go back through their email and scan past reports for relevant stories. Headquarters staff hope to be able to identify trends not just in the past, but to predict nascent crises. We asked the OPSCEN staff what they *wanted* field staff to do with the reports they compile, and they were unequivocal -- they want field staff to take action.

Related Work

This project presents the challenges of mining documents and text and working with spatial-temporal data. As summarized by Brehmer et. al.[1], there have been diverse approaches to analyzing large document collections and presenting visual clusters, including interactive tag clouds, hierarchical trees, and scatterplots. For example, the *Overview* tool[2] for investigative journalists allows users to hierarchically cluster text documents based on content similarity, visualizing the the collection as a tree. Our project has several similar characteristics: we will mine text to identify trends in current events. However, the project data is inherently structured, coming from news briefs selected by a team with specific goals and audience in mind, and we are able to pre-identify attributes for the users to cluster and search articles by, discussed in the Data section section below. Also, we are dealing with spatial-temporal data.

We also examined current crisis-mappers and related tools:

- LRA Crisis Tracker[3]:

This tool is a joint project of NGOs Invisible Children and Resolve, updated daily with recent activity by suspected fighters of the Lord's Republican Army (LRA), with the ability to filter by type of encounter and by date. This data is similar in that it is manually curated and added to the visualization, though their location data is much more granular. We found the date slider at the bottom of the map especially appealing.

- CrisisNet (Ushahidi team)[4]:

This is a in-development project that trawls Facebook and YouTube's APIs for geotagged posts from a pre-determined location (thus far, Syria only). Content is then automatically translated, categorized and geotagged to a specific region or city. The data is represented on a heatmap, split into regions of Syria. A tooltip displays the number of reports for the highlighted region, and clicking on a region will brings up an example post. This data is similar to ours in that it has to be auto-tagged and categorized with natural language processing, and that it is plotted on a heatmap split by region. However, our data is curated, not automatically gathered by API, and thus needs to be added by the user. Additionally, we want users to be able to view all news stories for a given area, and text-based data isn't well suited to being viewed exclusively in a tooltip.

- Geovisual Analysis and Crisis Management[5]:

---

[1] Overview: The Design, Adoption, and Analysis of a Visual Document Mining Tool For Investigative Journalists. Matthew Brehmer, Stephen Ingram, Jonathan Stray, and Tamara Munzner. IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis 2014).

[2] https://www.overviewproject.org/

[3] http://lracrisistracker.com/

[4] http://crisis.net/projects/syria-tracker/

[5] Geovisual Analytics and Crisis Management. Brian M. Tomaszewski, et. al. Proceedings of the 4th International ISCRAM Conference (Delft, the Netherlands), May 2007. http://www.personal.psu.edu/acr181/ISCRAM_2007_Accepted.pdf

The Pennsylvania State University team have created an interactive visualization tool that integrates mostly georeferenced data from multiple sources, including news bulletins, emails, photos, and video feeds, into one environment to improve situational awareness, decision-making, and crisis management throughout a crisis. The tool provides another example for how to visualize text data with a temporal-spatial component. However, the end goal and user is drastically different than our project's: the objective of the Geovisual Analytics application is to aid crisis managers throughout a crisis and response.

Data

OPSCEN provided 283 files of news briefs,[6] spanning 2013 to 2014. Each news brief is dated and provides a short one-line statement about a news item, as well as a paragraph-long summary and url to the source article. The news items are grouped by country and region. The attributes inherent in each news brief are: date, region, country, event title/one-liner, event summary, and source article url. These are all categorical datatypes, except for the date, which is quantitative. We will use all of these attributes in our project. Additionally, we plan to derive several other fields:

- event categories or "tags" (categorical): we plan to use a word count algorithm and latent Dirichlet allocation (LDA) to characterize each news item by topic to allow users to filter news items by item type. For instance, an item like "New malaria tool to help track insecticide resistance," may be tagged as "disease." Tags may or may not be hierarchical--for example, "flood," "earthquake," "storms," "tsunami," and "drought" may be grouped into a higher "natural disaster" category. If LDA does not work appropriately, we will use a top-down approach and pre-select categories and search for these categories and related words to cluster news items.
- Number of people (quantitative): many news items list the number of people affected by the event. We will attempt to pull out this number from each news summary to be able to categorize and filter news items by number of lives impacted. However, not all news summaries include this number; it also may be difficult to pull out, due to the presence of other written numbers and digits in a paragraph and the structure of a given sentence.
- Month and year (ordinal): we will break up each each date into more granular fields to be able to identify and show trends by day of the week, month, and year.
- quotes (categorial): the body of some news summaries include quotations. We will identify these quotes and might include them in the visualization.
- entities (categorical): We will use Name Entity Recognition to identify organizations, locations, and people and link news items to entities, allowing users to search and filter by entities.

Tasks/Questions

[6] **https://www.dropbox.com/sh/a9e62mraviveldt/AACcjVTmMPZtRqzByDL8n82ia?dl=0**

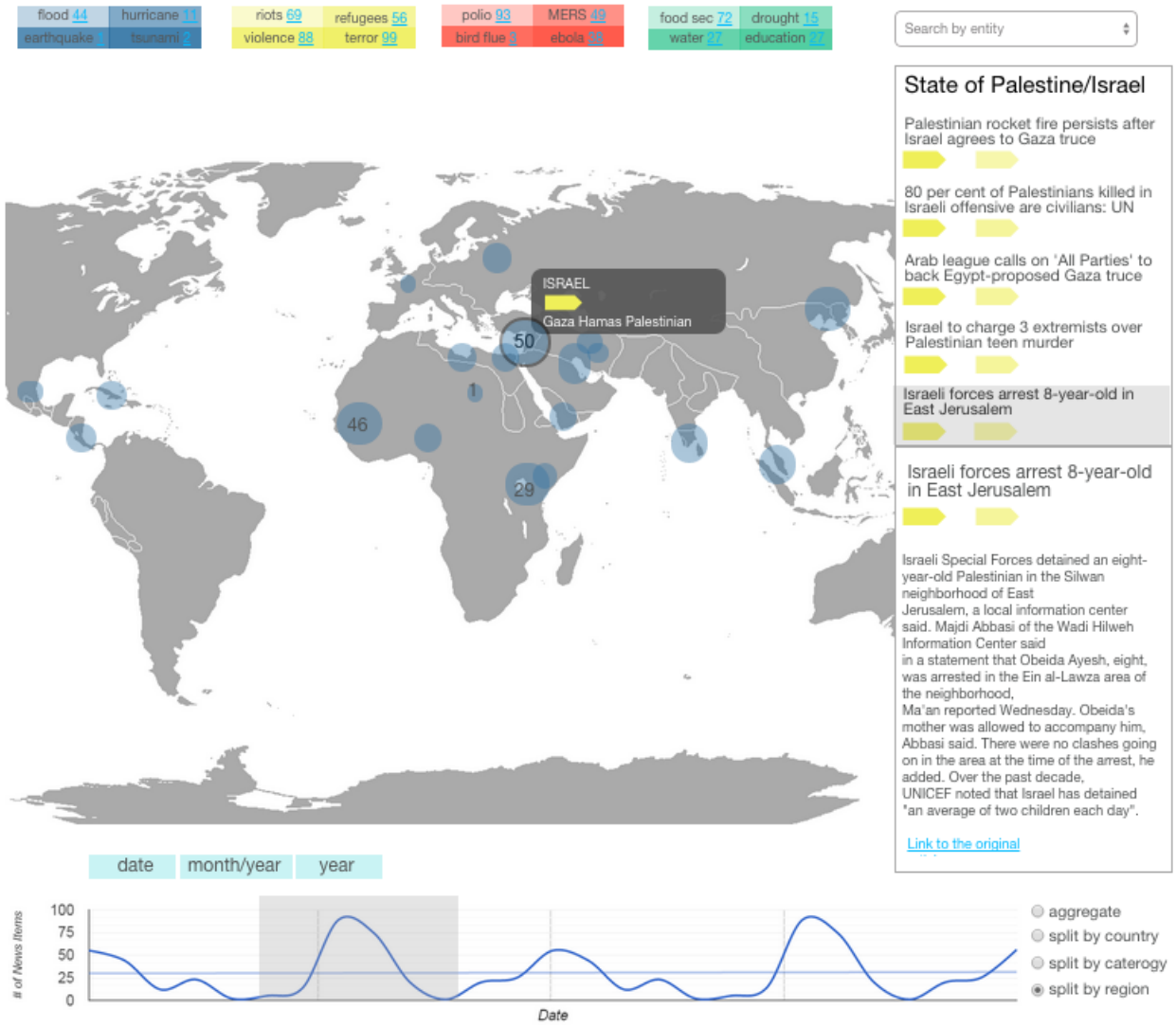
T1: What crises or security-related events happened today (or within a specified timeframe)?
We aim for users to be able to select a timeframe and be able to get a sense of the number and type of events that occurred in a country or region.

T2: What are long-term or nascent trends? How do regions or countries compare?
Our goal is to allow the user to identify spatial-temporal patterns. The user will be able to filter news items by location, category tags, and entities and see trends on the map;
the user will also be able to see news items in aggregate or split by the selected filters (country or category) in the accompanying line graph. We hope to allow the user to aggregate by either number of events/news items or people impacted. The filtering tools will allow users to compare two or three countries or category types.

Open Questions (so far):
We expect many questions and design problems to come up as we develop our visualization. Some questions we have currently identified are listed below:

1. How to cluster and handle hierarchical categories:as noted above, there may be certain categories that can be broken up into sub-categories. However, classifying news items into several levels based on text in the news briefs may be difficult due to the limited summaries. If we do have a hierarchical structure to the categories, we still need to determine how best to visualize this and the connections.
2. How to visualize new stories: if there is a sudden outburst of violence in a region or a serious disease outbreak, this is especially relevant information for UNICEF staff, and one of the things that staff particularly looks out for when browsing the data in its current formal. However, in the short-term, when the crisis is just beginning, this may show up in the heatmap as just one or two stories (and thus get drowned out by long-running conflicts and crises).
3. How to best represent number of stories by country: in the visualization below, we show the number of stories by circles. However, we recognize that circles may not be the best way to represent this information. For example, if there are a lot of stories for two small neighboring countries, the circles may overlap significantly, making it difficult to determine and compare the size of each. One option is to not use circles at all but represent number of stories by the color intensity of the color fill of a country.
4. How field staff currently use the dataset, and how they would *like* to use or visualize the news brief data. As described above we have spoken to staff at UNICEF headquarters but hope to speak to one or two field staff members to better understand how they currently consume and use the news briefs and the potential they see in the dataset.
5. How to handle the back-end storage of the data -- we don't know if we need a fully developed backend system, and if so how extensive it has to be. We would greatly appreciate technical guidance on how to handle our data.

Visualization Sketch:



Workplan

We plan to work collaboratively on all the phases of the projects, however, we assigned a responsible team member to each project phase mainly to coordinate the individual tasks and to make sure our work won't overlap.

Phase			Lead by
Data Processing	Document parsing		Ady
	Model direct attributes		Meredith

	Model derived attributes	LDA - topic modeling	Andrea
		NER - name entity recognition	Ady
	Database Development		Andrea
Revise Visualization Sketch			Group
#1 Client checkpoint			
1st prototype coding	Map		Andrea
	Filters		Meredith
	Timeline		Meredith
	Document Viewer		Ady
2st prototype coding	Map		Andrea
	Filters		Meredith
	Timeline		Meredith
	Document Viewer		Ady
#2 Client checkpoint			
Final version coding	Map		Andrea
	Filters		Meredith
	Timeline		Meredith
	Document Viewer		Ady
Project write-up			Group
Project demo video			Group
Project presentation to the client			