Visualizing News Report

This is a website that can give you a brief idea of the topics on which a specific news source posted most and fastest within a given time range.

Meng Yang (my1421@nyu.edu)

Ze He (zh700@nyu.edu)

Xiaoyan Wang (xw1108@nyu.edu)

Project page (on Github): https://github.com/miniwangdali/IVProject

Video: https://vimeo.com/196909864

Working demo: http://news.heiseric.com/

Description

The problem is that when a story is reported, we can always find the same story on several different news websites at different time. If we know which news website report this kind of story fastest, then next time, we can go directly to that website to read this type of story instead of wandering around and read second-hand news.

Knowing the pattern of news report is good for all customers who read news. And also for news press, if they know the pattern, they can work harder on those topics they reported too slow.

It is important for investors to obtain the latest information in a particular area as quickly as possible so that they can smell the tendency and make decisions or solutions towards events and incidences. An investor can take the lead if the one knows the website on which the fastest news are always posted in each specific field. Also, it is important to know whether the news is real or fake. And considering the development of the website, it is better to be able to guess which website will have the freshest news in the future. So, investors need the information about how fast the news posted in each website in each topic.

On the other hand, a website should know its competitors in its focusing topics and find chance to excel the leaders. The faster can gain a larger quantity of flow which means more benefits. So, the website need the information about posting news speed of itself and the competitors.

Questions

- The number of “first-mentioned” topics one news website has reported during a given time.
  According to the density of the rects in one card, it is easy to get the idea of how many first mentioned topics this website has reported.

- The distribution of the “first-mentioned” topics of one new website during a given time.
  The rects in a card is colored according to the topic it refers to. It’s easy to know the distribution of first mentioned topics with the colors’ distribution.

Data Abstraction

Attribute Name	Attribute Type	Meaning	Values	Derived? (if yes explain how)
article_id	Categorical	unique id per article	object identifier (OID)	No
article_url	Categorical	original link to the article	web url	No
event_groups_group (Topic groups)	Categorical	A major event	16 event groups	No
event_groups_type (Topics)	Categorical	A subsection of an event group	78 event Types	No
event_name (Related Corps)	Categorical	Name of the company	8,000-plus U.S. public equities	No
first_mention	Categorical	A story that has not been mentioned on the internet for at least two weeks.	true, false	No
harvested_at	Ordered	Date Accern received the article	yyyy-MMdd'T'HH:mm:ss.SSSZZ (UTC TIME ZONE)	No

Website	Categorical	Website name	Every website in the data	Derived from *article_url*

Our data comes from a set of news report data given by Accern Low-Latency Sentiment API[1]. The API gives data that shows all news appeared in the internet and records relative information of the news. The types of data are “Public News Websites”, “Public Blogs”, “Press Release”, “Twitter”, “Financial Document”, “Other Social Media (Tumblr)”. Our data is a pre-collected csv data and we do some cleaning and processing on the data.

Our preprocessed data is saved in Mongodb.

Related Works

1. Newsflow[2]

Newsflow is a dynamic, real-time map of news reporting, which displays both the latest top stories as well as the news organizations which covered them. All articles are from the last few minutes.

Viewing news in this way lets us see how the choice of 'top stories' by news bureaus is geographically unequal, or rather, what areas of the world are neglected by various national news sources.

The ability to view such data in real time offers viewers a chance to see how journalists shape national attention as stories unfolds.

Both this project and our project are focus on the relationship between stories and websites or organizations. But it focuses on the latest top stories as well as the news organizations, and our focus is the fastest stories as well as the news websites.

[1] https://www.accern.com/

[2] Website: http://newsflow.cartagen.org website down

2. Visual Analysis of News Streams with Article Threads[3]

This paper proposed a purely visual technique that permits to see the evolution of news in real-time. The technique permits to show the stream of news as they enter into the system as well as a series of important threads which are computed on the fly. By merging single articles into threads, the technique permits to offload the visualization and retain only the most relevant information. Such technique will be helpful in the design of news streaming visualization in our project.

3. The NewsStream Portal[4]

NewsStream is a web-based interactive news and blogs analytics portal. Prior to what happened, it focuses on the questions

- when something important happened
- what was the amplitude (volume)
- what was the related sentiment.

Its main functionality is the projection of relevant documents (news articles and blog posts) and their sentiment onto a timeline, summarization of the documents by aggregation, and drill-down to the original documents.

Design Iterations

At first, we gave out three completely different designs: heatmap, pie chart and line chart. And then we tried to combine it. We mix all the options and the characteristic in each design.

But during our process on website, we found that it was difficult to show all the data on the web page at once. Also it would bring the problem of performance for user to explore the web page. Then for more precise and more accurate visualization, we deleted the options for user to choose the companies they would be interesting in.

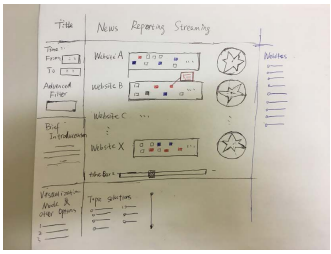
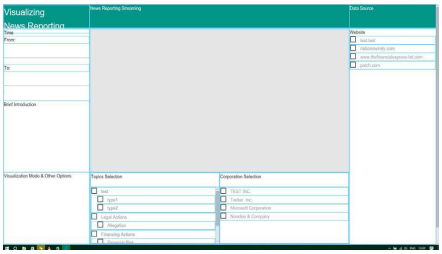
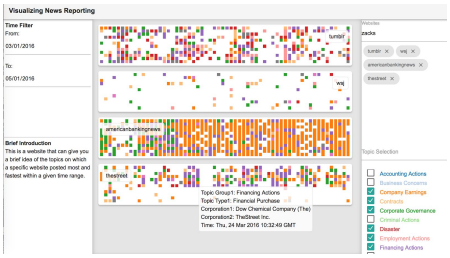
Later, We found that we could get more specific, clear and important information on topics' group rather than on topics' type. So we deleted the options on second level topics' type.

---

[3] Document: https://bib.dbvis.de/uploadedFiles/251.pdf

[4] Website: http://newsstream.ijs.si/ Document: https://arxiv.org/pdf/1508.00027.pdf

Also, user would actually focus on the websites they would like to observe and there are over 20 thousands websites. It is impossible for user to explore the whole selection list and select or deselect each websites. Thus, we brought in autocomplete feature for user to type in the name of the website and we give their a list of possible websites for selection. And we change the selected website list to a board of website chips which is more clearly on showing the selection and easier for user to modify their selection.

Final Visualization

Below the title row, the container is divided into three columns. The left column is a time filter, which will filter out all news data in a given start date and a given end date. The upper part of the right column, is a autocomplete search bar, when user finish a search operation, a card of this website will

show up in the middle column, and will be placed below all the existing card. The lower part of the right column, a list of topics will show up for users to choose. By default, all 18 topics are chosen, once user cancel one topic, the related rect will disappear from the website card.

The main component of this visualization is the middle column, where all the website cards exist. Each website card shows all the stories that are first mentioned by this website and each story is represented by a rect, colored by the topic group this story belongs to. All the rects are distributed by the post time. In the website card, the X-axis is divided into 60 points. Each one represents a time range as 1/60 of the overall time ranges. All the stories posted during this small time range will be placed from top to bottom on the same X-axis point, its Y-axis is decided by its post time during this small time range, earlier one will be placed closer to X-axis.

Findings

1. Which websites usually have the first hand report on the internet?

As we can see in the above screenshot, we can intuitively know that tumblr, zacks and thestreet have more first-mentioned news on their website which oppositely chinadaily and twitter have less. This means it is better to go through the tumblr, zacks and bloomberg if one would like to get a first-hand news or discover some news quickly.

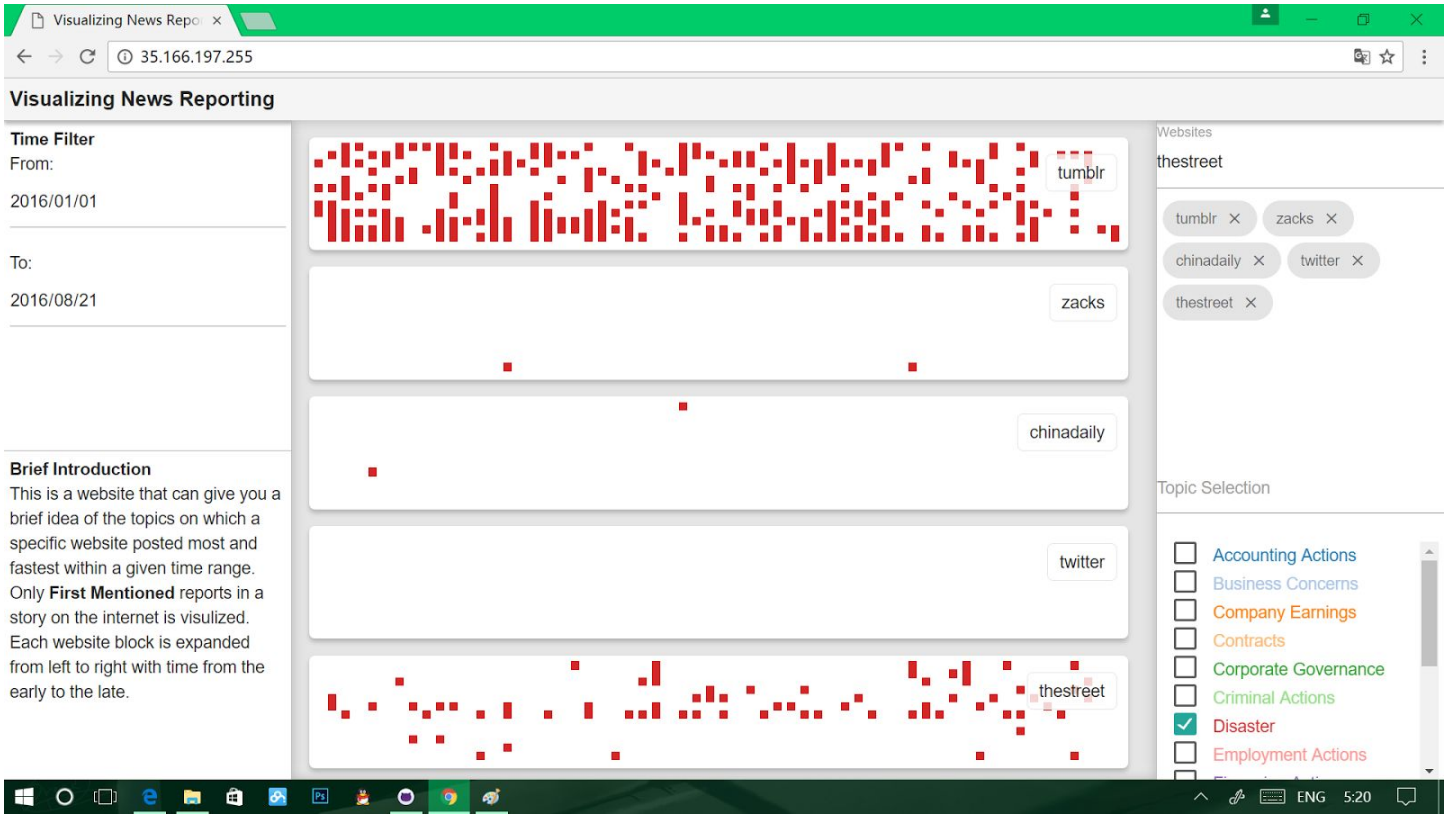
It is interesting that tumblr, as a blog website, has almost the largest amount of first-mentioned news but twitter, as a social network website, does not.

2. What kinds of topics does each website usually first mentioned?

As we can see in the above screenshot, we can clearly find that tumblr is good at reporting different kind of news. Zacks is mainly focus on Company Earnings news.

It is also interesting that tumblr has a good ability on reporting disaster news. See next question.

3. For the topics of disaster, which websites usually have the first reports?



As we can see in this screenshot, tumblr and thestreet usually have the first report for disasters. This result conveys that tumblr, as a blog website, is also a good place to get first-hand information about disaster happens around the world.

Besides above findings, there are a lot of information waiting for us to dig out. Above findings are just some examples for showing how useful and easy it is that our website could express and extract information among a lot of news data. We hope it would be helpful to the researcher and investigator in the future.

Limitations and Future Works

One problem we have not successfully solve is the overlapping of news articles. Since for some most influential websites, there are too many first mentioned articles released in a short period, however the space on screen is limited, these articles can't fit in together if they are published at the same time. This kind of overlapping of news reports will cause lose of information and might lead to some wrong visualizations in worst case.

Another limitation is the latency of autocompletion in websites search. This is because the large amount of websites in our data, a total of over 20000 unique websites.

Future works could be made in optimizing the performance of this project under large and real time streaming data flow. As well as improving the interaction with article blocks to overcome the problem of news overlappings.