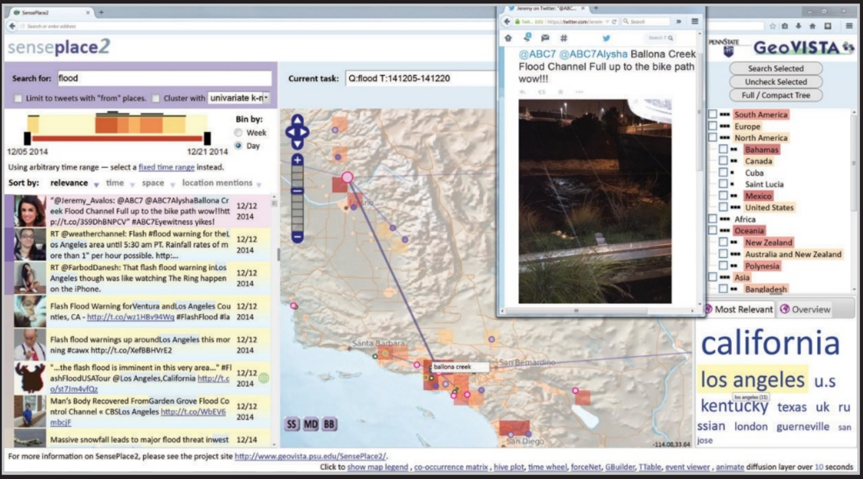
PENNSTATE 1855
GeoVISTA Center

Building the compute-power to make sense from 130 million Tweets

Monday, 22 December 2014

GeoVISTA Center researchers are leveraging Twitter to make sense of crisis situations. When hurricanes, disease outbreaks, fires, or other crisis events happen, social media react. This reaction can supply crisis managers with important information, but that information is hidden in a mass of noise. To extract the signal from the noise, Penn State's GeoVISTA Center has developed SensePlace2, a web-accessible geovisual analytics tool that collects, analyzes, and visualizes millions of tweets. The information foraging and sensemaking tools in SensePlace2 allow users to explore tweets through ad hoc queries about key topics of interest. Now, an innovative collaboration between the GeoVISTA Center and the Penn State Institute for CyberScience (ICS) harnesses the power of Apache Hadoop and cluster computing to enable interactive analysis of spatio-temporal trends in hundreds of millions of public social media posts.

The SensePlace2 user interface displays results for simple or compound searches and supports overview and detail maps of tweets, place-time-attribute filtering, and analysis of changing issues and perspectives over time and across space as reflected in tweets. As Alexander Savelyev, lead developer for the SensePlace2 interface, explains, "Dynamic linking among views allows an analyst to highlight a place on the map and quickly see the tweets mentioning that place, those posted from that place, and any other places connected to that place through the tweets or the users who posted them." For example, in a search for tweets most relevant to the recent flooding event in California shown below, those mentioning Los Angeles were selected by clicking on the place in the tag cloud; from the filtered tweets a link is followed to an image of an impending bike path flood.


senseplace2

Search for: flood

□ Limit to tweets with "from" places. □ Cluster with univariate k-n

Current task: Q:flood T:141205-141220

Bin by: ○ Week ● Day

12/05 2014 12/21 2014

Using arbitrary time range — select a fixed time range instead.

Sort by: relevance time space location mentions

"@Jeremy_Avalos: @ABC7 @ABC7AlyshaBallona Creek Flood Channel Full up to the bike path wow!!http://t.co/359DhBNPCV" #ABC7Eyewitness yikes! 12/12 2014

RT @weatherchannel: Flash #flood warning for theLos Angeles area until 5:30 am PT. Rainfall rates of more than 1" per hour possible. http:... 12/12 2014

RT @FarbodDanesh: That flash flood warning inLos Angeles though was like watching The Ring happen on the iPhone. 12/12 2014

Flash Flood Warning forVentura andLos Angeles Counties, CA - http://t.co/wz1HBv94Wq #FlashFlood #la 12/12 2014

Flash flood warnings up aroundLos Angeles this morning #cawx http://t.co/XefBBHVrE2 12/12 2014

"...the flash flood is imminent in this very area..." #FlashFloodUSATour @Los Angeles,California http://t.co/st7Jm4vfQz 12/12 2014

Man's Body Recovered FromGarden Grove Flood Control Channel « CBSLos Angeles http://t.co/WbEV6mbcjF 12/12 2014

Massive snowfall leads to major flood threat inwest 12/14

@ABC7 @ABC7Alysha Ballona Creek Flood Channel Full up to the bike path wow!!!

Search Selected
Uncheck Selected
Full / Compact Tree

□ South America
□ Europe
□ North America
□ Bahamas
□ Canada
□ Cuba
□ Saint Lucia
□ Mexico
□ United States
□ Africa
□ Oceania
□ New Zealand
□ Australia and New Zealand
□ Polynesia
□ Asia
□ Bangladesh

Most Relevant Overview

california los angeles u.s kentucky texas uk russian london guerneville san jose

For more information on SensePlace2, please see the project site http://www.geovista.psu.edu/SensePlace2/.
Click to show map legend, co-occurrence matrix, hive plot, time wheel, forceNet, GBuilder, TTable, event viewer, animate diffusion layer over 10 seconds

SensePlace2 user interface showing tweets relevant to recent California floods

SITE SEARCH
Search GeoVISTA Go
Or Search: Penn State, People, Departments

NEWS ADMIN
Submit News
Edit News

FEATURED VIDEOS
Flu Data Analysis with GeoViz Toolkit
Routesketcher
Health GeoJunction
Demonstration of SensePlace
GeoVISTA CrimeViz
Geo-historical Context Support in the Context Discovery Application
See more videos . .

GRAD HIGHLIGHTS
Kevin Ross
Adrienne Gruver
Jin Chen
Craig McCabe
Tom Auer
See all grad highlights . .

Scott Pezanowski and Dr. Frank Hardisty lead the GeoVISTA Center effort to scale up basic SensePlace2 visualization and query capabilities to cope with hundreds of millions of tweets, with more streaming in every day. Built upon Apache Solr, an open-source search platform, SensePlace2 can analyze over 130 million Solr Twitter documents that not only contain tweet text, but also metadata including when and where the tweet was posted, the locations mentioned, the Twitter ID of the poster, whether the tweet is a reply or a retweet, and more.

To support flexible queries that facilitate crisis situational awareness and other possible applications, SensePlace2 requires efficient data storage and a computer infrastructure that enables quick retrieval and display. In early versions of SensePlace2, a simple search took a fraction of a second. However, as the index grew larger, some compound searches began to take as a minute. For a web application that provides user interaction, this is unacceptable," Pezanowski explains.

The need for better performance prompted Pezanowski and Hardisty to collaborate with Chuck Gilbert and Pierre-Yves Taunay of ICS. The team has experimented with an ICS server cluster that runs the Cloudera Express distribution of the Apache Hadoop software, which includes SolrCloud for distributing Solr across multiple machines. The machine cluster is comprised of an edge node and four data nodes, each of which has Dual 8 Core processors, 256GB of RAM, 1.2TB 10K SAS drives, and a 10G Ethernet connection. These computational resources achieve a substantially better search performance.

Pezanowski notes, "Initial comparisons confirm the ICS SolrCloud cluster provides not only significantly improved SensePlace2 search times over a GeoVISTA server running a single instance of Solr, but also much more consistent search times." Return times for a typical SensePlace2 search, which uses a combination of search criteria, sorting, and multiple different aggregate counts of matching documents, have decreased from nine seconds to only two seconds in the SolrCloud cluster. The most intensive searches use a geospatial sorting. For example, a SensePlace2 user may be interested in tweets that have location mentions near a point he or she chooses on a map of the world. Solr searches for documents moving away from this point until the desired number is returned. This kind of search, which used to require 20 seconds or more, now takes seven seconds on the SolrCloud cluster.

Pezanowski anticipates further collaboration with the ICS team as they improve and scale their hardware and software for faster response times. "In today's world, searches need response times much faster than what we are currently achieving to meet increasing user expectations."

Dr. Alan MacEachren, Director of the GeoVISTA Center, adds, "Our next goals for SensePlace2 are to integrate sources of information beyond Twitter such as Instagram or news feeds and to add support for analyzing information diffusion over time in order to understand how the public reacts to official crisis alerts and warnings as well as to learn about place-related human behavior more generally. Achieving this will put even more pressure on the team to develop innovative strategies to support not only much larger data volumes but complex, on-the-fly analysis."

Scott Pezanowski is a Senior Research Analyst at the GeoVISTA Center and focuses on geospatial web application development. Dr. Frank Hardisty is an Assistant Director of the GeoVISTA Center, focusing on applying computing power to spatio-temporal trend detection. Alexander Savelyev is a Ph.D. candidate in Geography and lead developer for the SensePlace2 user interface. Chuck Gilbert is the Sytems Architect and Systems Team Lead for Advanced CyberInfrastructure at the Institute for CyberScience.

Pierre-Yves Taunay is currently a PhD student in Mechanical and Aerospace Engineering at Princeton University.

Dr. Alan MacEachren is Director of the GeoVISTA Center and leads its efforts in visual analytics and geographical information retrieval.

For more information, contact **Krista Kahler** at kck12@psu.edu
Visit http://www.geovista.psu.edu/SensePlace2/ to learn more.

Questions/ Problems with this site? Contact the WEBMASTER | Adapted from a design by FREE CSS TEMPLATES.