Threat News Explorer Technical Description

Introduction

The Threat News Explorer (TNE) is a web application created by the Western Wildland Environmental Threat Assessment Center (WWETAC). TNE facilitates assessment of wildland threats by collecting and displaying news articles on the web that discuss these threats. The user can view articles for specific predefined threats such as bark beetle, climate change, etc. Articles are saved in an archive database for analysis purposes. The application is hosted in the Amazon Web Services cloud and is available at http://wildfireapps.org/TNE/TNE.aspx.

The Google News Search API is used to retrieve news articles on a daily basis for predefined threats. The articles come from various news reporting web sites around the world and thus provide a method of collecting information with a wide reach. By retrieving, analyzing, and storing news articles, the Threat News Explorer can help identify trends as expressed in the news.

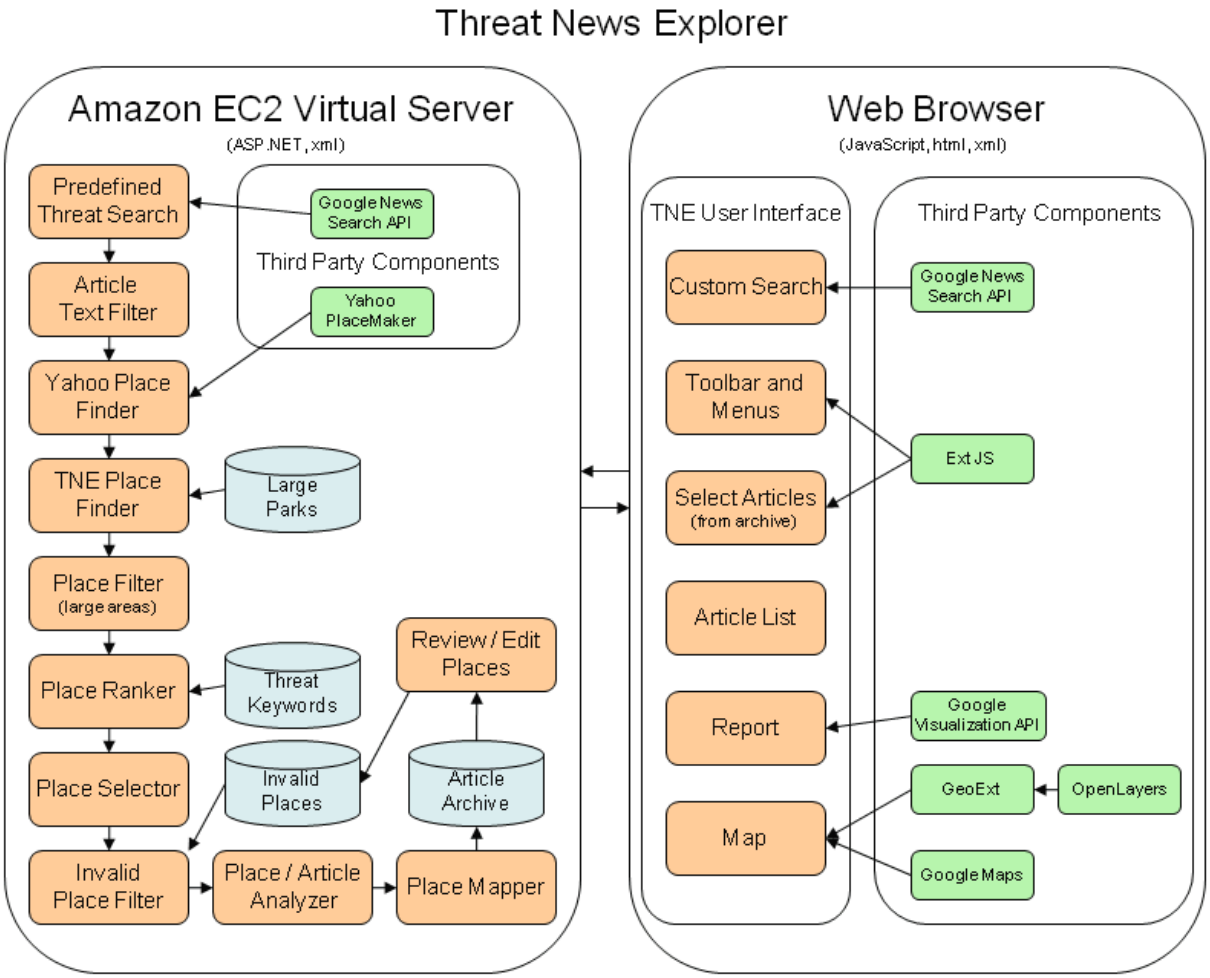

Figure 1. Threat News Explorer application design.

Article Processing, Storage, and Retrieval

Articles are retrieved from the Google News search on a daily basis. The articles are compared with existing articles in the archive database to avoid duplication. Any new articles are added to the database sorted by date in descending order. A copy of the web site is stored for internal analysis. The data is stored in xml format. The following information is stored for each article:

1. Threat Name – The name of the threat defined internally by TNE.
2. Title – The title of the article as returned by the Google search results.
3. URL – The URL for the web site containing the article
4. Published Date – The published date.
5. Content – A short snippet of the article provided by the Google search results.
6. Run Date – The date the Google search was conducted.
7. Archive Filename – The filename for the copy of the web site stored internally by TNE. This copy is not provided to the general public but is used for internal analysis. If TNE was unable to obtain a copy, the error message returned from the web site server is stored instead of the filename.

In addition, the archive database stores a list of places found in the article. This list only includes the places that are ranked highest for their relevance to the particular threat. The following information is stored for each place:

1. WOEId (Where on Earth IDentifier) – This is a unique reference 32-bit identifier assigned by the Yahoo! GeoPlanet project to identify any feature on Earth.
2. Place Name – This is the name of the place as returned by the Yahoo PlaceMaker service.
3. Latitude – The latitude of the place.
4. Longitude – The longitude of the place.
5. Reference Text – The Yahoo PlaceMaker service returns the character index where the place name was found in the web page text. Threat News Explorer uses that information to parse out the entire sentence containing the place name. If the place was mentioned more than once in the article, there will be multiple sentences stored in the reference text.
6. Rank – The rank is calculated by TNE based on the PlaceMaker confidence level, the total length of all the sentences in the reference text, and the relevance of the sentences to the wildland threat. The rank is used to select which places will be stored with the article in the database.

The application will initially display just the most recent articles. The Select tab to the left of the map allows the user to view all the articles in the archive or selected articles. The user can specify a published date range to select articles within that range. The user can also specify a search term and whether to search for the term in the article title, publisher, summary, or the entire article. By default, only articles that mention places within the current map extent are displayed. The list of articles changes as the map extent changes from pan and zoom actions. A checkbox on the Select tab allows the user to override this behavior so that all articles are displayed regardless of the map

extent. Articles that are not associated with any places will only be displayed when the map is zoomed out.

Mapping News Article Topics

The TNE application includes a component which maps places mentioned in the articles that are relevant to the threat being discussed. This map will help to identify areas where the threat is of particular concern as expressed in the news. This process involves custom software in addition to open source mapping components such as OpenLayers and GeoExt, a Google map layer, and a Yahoo service called PlaceMaker. The following steps are taken to create the map:

1. Collect news articles that discuss the predefined threat.
2. Filter the articles to remove html and preserve the text.
3. Send article text to PlaceMaker and receive a list of places found in the text.
4. Send article text to an internal place finder engine to find important places that may be missed by PlaceMaker.
5. Eliminate places that cover a large area such as continents, countries, or states.
6. Rank the places by their relevance to the threat being discussed.
7. Select only the highest ranking places and remove the rest.
8. Remove places that are known to be invalid.
9. Combine information from multiple articles referring to the same place.
10. Add the places to the map using point markers.
11. Review and remove invalid places.

The first step uses the Google News Search API to create a database of news articles for specific predefined threats. Each threat is searched on a daily basis and the results stored in a database on the TNE server. The server makes a copy of each news web page for internal use. No attempt is made to drill down into additional web pages that may be referenced in HTML links.

Next, the articles are sent through a filter which strips out HTML tags and scripts, and preserves the article text. This filter will also eliminate text that is not in sentence form in order to extract just the article discussion and avoid miscellaneous links.

The filtered article text is then sent to the Yahoo PlaceMaker web service. Since the process of identifying places from web page text can be error prone, this service ranks all places found with a level of confidence that the place was correctly identified. The service allows users to specify the level of confidence they are willing to accept for places that are returned. The TNE application uses a high level of confidence when calling the PlaceMaker service. The service returns an xml file to TNE with a list of places found. This list includes the name of the place, latitude and longitude coordinates, the confidence level, and where the place name occurred in the article.

Initial testing of the PlaceMaker service has found that important places such as parks, national forests, and other known wildland areas, are sometimes missed. Since TNE focuses on wildland threats, it is important to not miss these places when they are mentioned in the text. A GIS polygon

dataset was downloaded and processed for large wildland areas. A latitude and longitude was identified for the centroid of each polygon to obtain a point that can be added to the map. TNE will search for each place name identified in this GIS dataset. Any places that are found will be added to the list of places already found by the Yahoo PlaceMaker service.

The Yahoo PlaceMaker service will identify places that cover large areas such as continents, countries, or states. Since we are mapping points and not polygons, these large places are eliminated to avoid confusion.

Each place that is found is ranked by its relevance to the threat being discussed. The ranking algorithm includes consideration of the PlaceMaker confidence level, the length of the sentence where the place was mentioned, and keywords found in the sentence. Testing has indicated there are sometimes very short sentences that mention places. These are usually not a part of the main article discussion, but instead represent links to other web pages. By reading representative articles, a list of keywords was created for each threat. These keywords are words that are likely to be used in a discussion of the threat. For example, the words “Bark”, “Beetle”, “Epidemic”, “Outbreak”, “Damage”, “Pine”, “Removal”, “Infest”, “Life cycle”, “Ravaging”, “Pest”, “Insect”, and “Ips typographus” are just some of the keywords identified for the Bark Beetle threat. Each keyword is assigned a number of points according to how likely its presence in the sentence indicates the threat is being discussed in that sentence. These extra points are added to the place rank for each keyword found in the same sentence as the place name. This system will effectively provide a higher ranking when place names are found in long sentences with many keywords, and a lower ranking when place names are found in short sentences with few or no keywords.

After assigning a rank to each place, the application will eliminate low ranking places. A rank lower threshold is defined, and places with ranks below that threshold are removed from the list. A maximum of 5 places are allowed for each article. To minimize articles without any places identified, at least one place is kept for each article regardless of rank. There may be some articles with no places identified.

The PlaceMaker web service sometimes makes mistakes in identifying places. For example, with the Bark Beetle threat the places “Pines, Steamboat Springs, CO, US” and “Pine Bark, Waleska, GA, US” are often returned. The TNE application provides an advanced user interface to identify these places as invalid so they will be eliminated from the list of places returned by PlaceMaker. This process is performed in an iterative fashion, where the advanced user periodically reviews new articles and identifies invalid places, and then performs an update to remove them from the database. The application also provides a mechanism to remove certain places from specific articles that are deemed invalid by the user.

The map has a Google base layer and an OpenLayers vector layer displaying point markers for the places identified in the articles. When multiple articles refer to the same place, the information from the each article is combined and stored with the place displayed in the map. The user can click on a place in the map and view a popup window displaying information for that place. The popup

window displays the title and link for each article and the sentences where the place name was found. As the user changes the map extent through pan and zoom tools, the list of articles displayed on the left will also change to include only articles that mention places within the map extent.

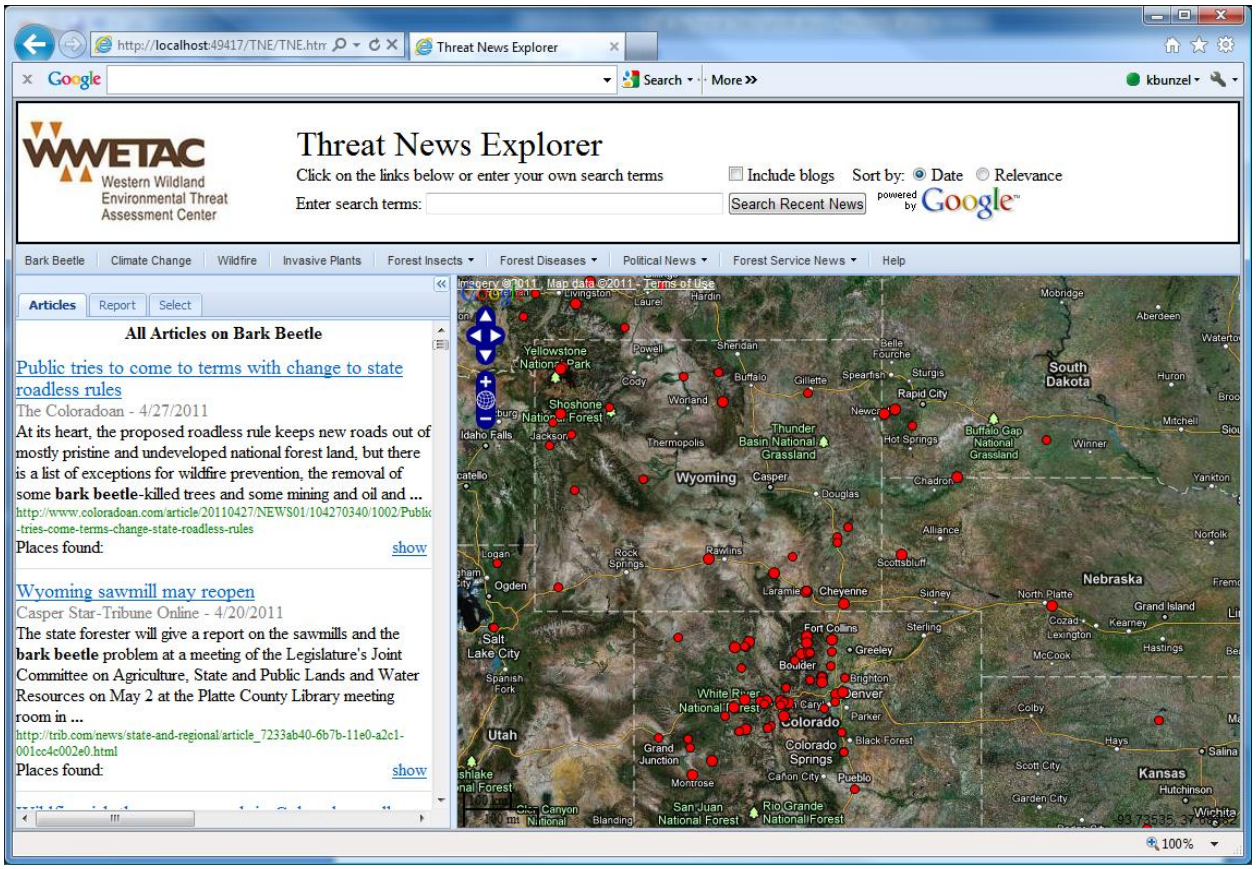

Figure 2. Threat News Explorer browser interface.