New Methods for Representing and Interacting with Qualitative Geographic Information

Contract #: W9132V-11-P-0010

Contract Period: May 23, 2011 – October 31, 2012

Principal Investigators:

Dr. Alan M. MacEachren, GeoVISTA Center, Penn State University

Dr. Anthony Robinson, GeoVISTA Center, Penn State University

Report on Component 2: Component 2 – Designing New Methods for Visualizing Text in Spatial Contexts

Alexander Savelyev, Scott Pezanowski, Anthony C. Robinson, and Alan M. MacEachren

<savelyev, spezanowski, arobinson, maceachren>@psu.edu

GeoVISTA Center, Department of Geography, The Pennsylvania State University

Submitted, Oct. 31, 2012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
04-09-2012	Final	May 23, 2011 – Sept. 4, 2012

4. TITLE AND SUBTITLE

Report on Component 2: Designing New Methods for Visualizing Text in Spatial Contexts

5a. CONTRACT NUMBER:
W9132V-11-P-0010

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Alexander Savelyev, Scott Pezanowski, Anthony C. Robinson, and Alan M. MacEachren

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

PENNSYLVANIA STATE UNIVERSITY , THE
408 OLD MAIN
UNIVERSITY PARK PA 16802-1505

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

US Army Engineer Research and Development Center (ERDC)
Topographic Engineering Center (TEC)
7701 Telegraph Road
Alexandria, VA 22135-3864

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; Distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Component 1 performed a broad review of existing visualization approaches to depicting qualitative spatial data, particularly that derived from text sources. These included both representation methods as well as interaction types that have had demonstrated utility in tasks that require the visualization of textual media that may originate from a variety of sources and may include a focus on the structure of documents, their relationships, and their associated entities. Using the understanding obtained during work on Component 1, and follow up consultation with USACE, we designed new methods that can be used to provide text information integrated with interactive maps. As previously discussed, we have focused our effort on two types of visualization integration: (a) representations that are overlaid on the map itself and (b) coordinated view approaches that offer interactive, visual coupling between separate text and map visualizations. SensePlace2 – an in-house geovisual analytics tool – was chosen as the testing platform for our design ideas, which were implemented as extensions and modules that fit into the original SensePlace2 data processing and visualization pipeline. This report outlines the key capabilities of the modules that were added to SensePlace2 as part of work on Component 2 and provides a detailed description of each of the major visualization components.

15. SUBJECT TERMS

geovisualization, visual analytics, social media, microblogs, cartography, qualitative geographic information, text analytics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE	SAR	14	19b. TELEPHONE NUMBER *(include area code)*
SAR	SAR	SAR			

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Component 2 – Designing New Methods for Visualizing Text in Spatial Contexts

Alexander Savelyev, Scott Pezanowski, Anthony Robinson and Alan M. MacEachren
GeoVISTA Center, Department of Geography, The Pennsylvania State University
<savelyev, spezanowski, arobinson, maceachren>@psu.edu

1 Introduction

Component 1 performed a broad review of existing visualization approaches to depicting qualitative spatial data, particularly that derived from text sources. These included both representation methods as well as interaction types that have had demonstrated utility in tasks that require the visualization of textual media that may originate from a variety of sources and may include a focus on the structure of documents, their relationships, and their associated entities. Using the understanding obtained during work on Component 1, and follow up consultation with USACE, we began designing new methods that can be used to provide text information integrated with interactive maps. As previously discussed, we have focused our effort on two types of visualization integration:

1. Representations that are overlaid on the map itself and
2. Coordinated view approaches that offer interactive, visual coupling between separate text and map visualizations.

The reason for this dual approach comes from the fact that most existing methods for qualitative data visualization have been developed for non-geographic applications and cannot be easily ported to an explicitly spatial analysis environment. Thus, an effective combination of spatial and a-spatial visualization techniques is required to address the unique characteristics of place-relevant text information.

SensePlace2 – an in-house geovisual analytics tool – was chosen as the testing platform for our design ideas, which were implemented as extensions and modules that fit into the original SensePlace2 data processing and visualization pipeline. This report outlines the key capabilities of the modules that were added to SensePlace2 as part of work on Component 2 and provides a detailed description of each of the major visualization components. In addition to this report, SensePlace2 also has a built-in legend, accessible through a link at the bottom right corner of the project's web interface, which provides an abridged summary of the information provided below. A video overview has also been completed as an additional deliverable within Component 2 and it illustrates much of the functionality described below.

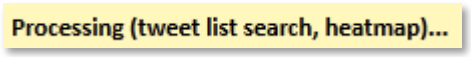
2 Access and Performance

The current version of SensePlace2 can be accessed using the following web address:

http://www.geovista.psu.edu/SensePlace2/app/

In order to prevent unauthorized access to the system, user authentication is required. When prompted, enter the username and password you have been provided with.

Interaction with SensePlace2 is primarily query driven. That is, users need to provide one or more search terms before most of the User Interface (UI) components can be used. In order to keep the user posted about the progress of the latest query, a status message is displayed at the top of the screen. Some of the status messages are directed at SensePlace2 users, while others are meant for the development team and can be somewhat cryptic. We are currently working on building a set of status messages that would be best fit for the general users. A typical status message would look roughly like this:



Once the query is complete, the status message disappears, the tweet list is populated with matching tweets and point symbols on the map appear. This set of actions indicates that it is now possible to interact with the display or initiate a new query.

System performance is, in general, dependent on the number of matches a given query has in the database. Thus, queries that use popular search terms (e.g. “fire” or “protest”) will take longer than queries based on less common keywords. Currently, some of the more popular queries can take as long as 1 or 2 minutes to complete due to the millions of records being processed (the research team is actively working on optimization of performance under separate funding, thus improvement can be anticipated over time). SensePlace2 caches query results in an aggressive fashion, which means that identical queries will return results much faster on the second run than on the first.

Although the SensePlace2 interface is regularly tested for consistency and stability, a 

button has been provided in case the UI experiences a catastrophic failure. Using the reset button will preserve the changes you have made to the UI during the current analytical session, and will likely fix all of the outstanding problems. If all else fails, the “Reload” button in your browser can be used to restart the UI. Under rare circumstances, the entire system might be down. Much progress was made to reduce such downtime but more work is ongoing.

3 Search Controls

A sample query based on terms “flee” and “Syria” is used below to demonstrate the functionality of SensePlace2 UI. A screenshot of the entire web interface, taken upon the completion of the query, is shown in Figure 1 below.

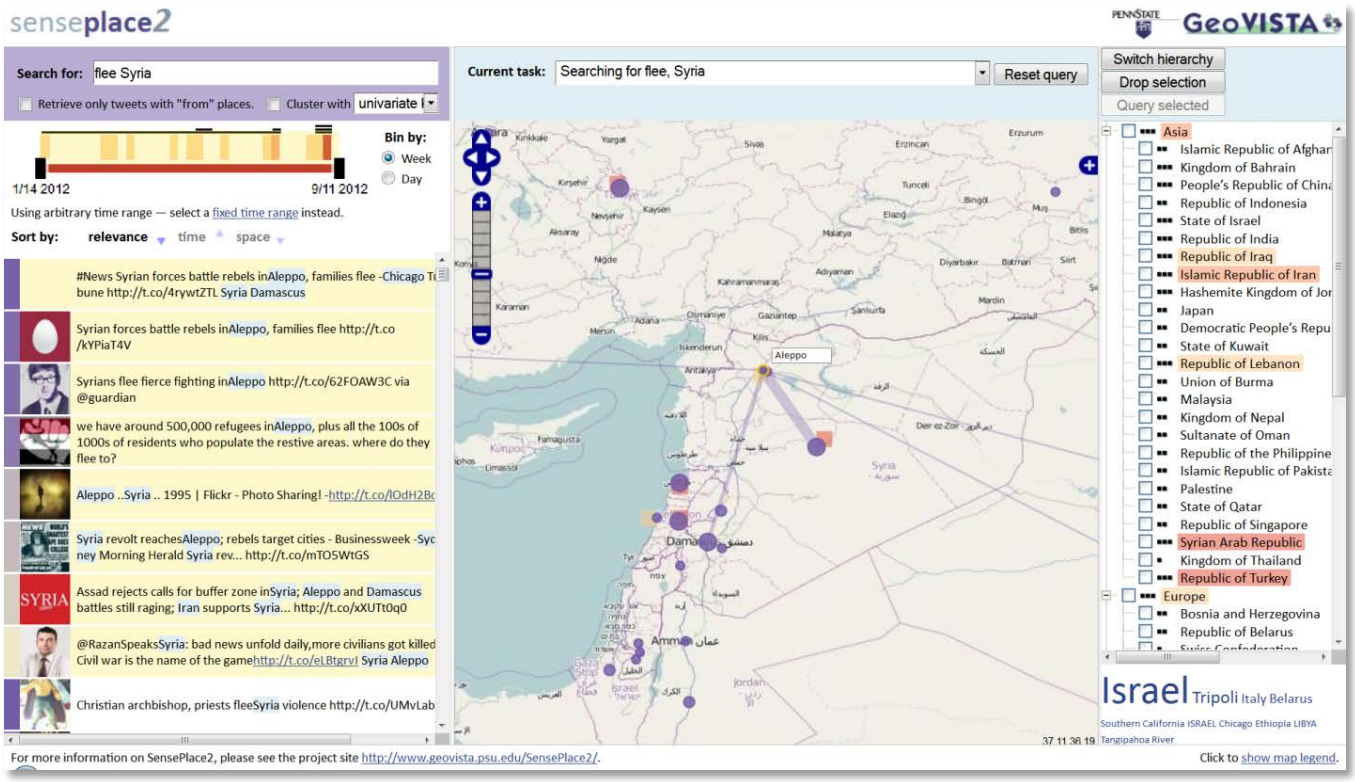
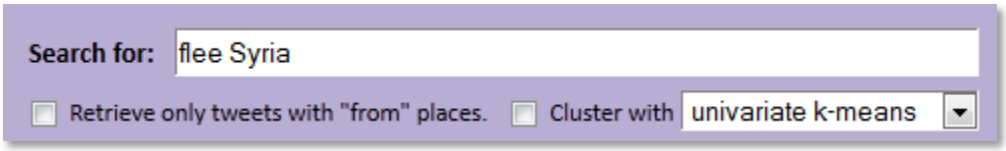


Figure 1, Sample SensePlace2 query.

Search controls are located in the purple zone at the top left corner of the web interface (shown enlarged in a figure below). Search controls can be used in three different ways.



3.1 Free-text query

As mentioned above, the current version of SensePlace2 is driven primarily by user queries. The "Search for:" input field allows users to insert one or more query terms of interest. You may currently search for single- or multi-word phrases (e.g. *football riots*), as well as an exact phrase by using double-quotes (e.g. *"football riots"*).

It might be possible to have some apparent mismatches between the data shown in different parts of the UI. For example, the heatmap may plot a single square on the map indicating a query match, while the tweetlist will show no results. This can be due to multiple reasons. A primary reason is that the heatmap reflects all matching tweets in the database while the tweet list shows only the 1000 most relevant tweets. In addition, we process queries using a combination of relational database search tools (PostgreSQL with PostGIS extensions) with a high-performance text search engine (Apache Lucene). This combination allows us to run sophisticated queries in near-to real time, yet sometimes results are slight mismatches as described above. Mismatch due to this process is minimal and only becomes apparent when few to none query matches are found. A planned upgrade

to the system is to index by location as well as time and concept; this will improve the match between the spatial aggregation shown in the heatmap and the other text-based displays.

3.2 Working with "from" places

One of the main features of SensePlace2 is that we extract geographic information from two sources. The first source is the *body of the tweet itself* (i.e. the names of locations that are mentioned in the tweet). We refer to information coming from this source as *"about" locations*, as people talk "about" them. The second source is the metadata associated with the tweet that often has explicit geographic coordinates in the form of latitude and longitude. We use the term *"from" locations* to refer to this kind of information, as people send tweets "from" them. "From" tweets are distinguished visually on the map by being shown as green rather than purple circles.

The number of tweets that have "from" locations is quite small (typically about 1.5% of all tweets, somewhat higher in crisis situations), and they tend to be drowned in the stream of relevant tweets with locations of the "about" kind. The checkbox labeled **"**Retrieve only tweets with "from" places" enables users to only bring back the tweets with associated "from" locations.

3.3 Clustering similar tweets

The **"Cluster with"** option allows users to apply one of two text clustering tools that group tweets into a small number of clusters. The resulting clusters are shown at the bottom of the tweet list using a few frequent terms that occur in tweets within the cluster. Clicking on a cluster in this display will bring the tweets from that cluster to the top of the tweet list. Clicking again will return to the default list.

3.4 Spatial Search

Spatial Search allows the user to find tweets related to a particular location on the map. Two kinds of spatial query are currently supported. The first is designed to retrieve tweets based on their proximity to an arbitrary location on the map, whereas the second is designed to retrieve tweets that make explicit references to a specific place-name.

3.4.1 Proximity to an arbitrary location

As part of this query, tweets are ranked as more or less relevant based on the distance between the "about" locations they contain to the arbitrary point specified on the map. Spatial Search will start at the point clicked and move away to find 1000 tweets with locations nearest the click point. An interesting visualization aspect of this search is that other locations mentioned in the top 1000 tweets will also be plotted on the map allowing the user to see other locations that are connected to the place of interest. This kind of query is activated by holding down the "Alt" key while clicking on the map.

3.4.2 Explicit reference to a place name

As part of this query, the user is allowed to restrict search results to those that mention a particular location or any location within a certain region. We are currently using the GeoNames hierarchy as our repository of hierarchical place names, although this functionality can be expanded to arbitrary geographies. For example, tweets can be filtered based on the congressional district they belong to in order to monitor election-related topics. This kind of query is activated through using the Place-tree, as described in Section 4.3 and in more detail in Section 8.

4 Overview and Detail

One of the principles for the analysis of the multi-scale data that gained the most prominence in the geovisualization and information visualization communities is the Information Seeking Mantra which is formulated as “Analyze first, show the important, zoom, filter and analyze further, details on demand” (Cockburn, Karlson, & Bederson, 2008; Shneiderman, 1996). SensePlace2 provides both overview and detail depictions of the Twitter data in the timeline, map, and place-tree views that are tightly coupled to the current observation scale and query parameters, as described below.

4.1 Timeline

The timeline displays the changes in the density of tweets over time and matches the parameters of the user query. Color shaded bands (overview) represent the number of matches the given query had in the entire database, with dark red indicating the time span with highest number of tweets. This time overview (week or day) is calculated on-the-fly, based upon user search parameters. The quintile category the count falls within is returned. The stacked black bars (detail) represent the number of query matches in the list of top 1000 relevant tweets.

Both color bands and the stacked bars use quantile-based classification scheme (quintile and tertile, respectively). By default, the width of the individual color bands is set to one week.

4.2 Map

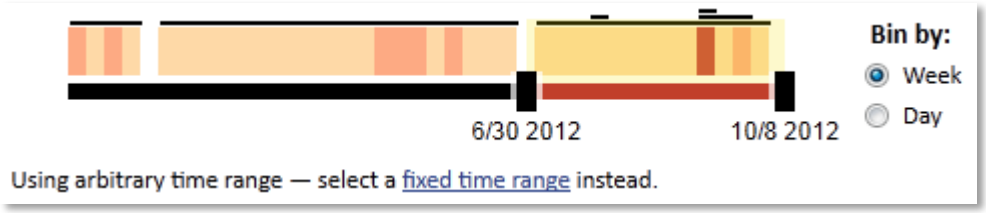
The map (as shown in Figure 1 above) uses a combination of heatmap and graduated point symbol displays. The heatmap (overview) displays the spatial density of tweets that match the term, time and place parameters of the user query using a quantile-based sequential color scheme. The heatmap calculations are performed through an on-the-fly database query that returns the quintile category the aggregate count of the cell falls within. Tweet density is calculated using the entire database. The top 1000 relevant tweets are plotted on top of the heatmap using graduated point symbols (detail). Tweets “from” and “about” a particular location are shown as purple and green, accordingly. The size of the point symbols represents the number of relevant tweets referring to that location, while their color density represents the aggregate relevance ranking of those tweets.

4.3 Place-tree

The place-tree highlights the locations that have been mentioned in the query results in a more structured fashion. Each of the nodes in the hierarchy is colored according to the number of matches the given query had in the entire database (overview), whereas the stacked black dots represent the number of matches in the top 1000 tweets (detail). Similar to the timeline and map overview, the hierarchy overview calculations are performed on-the-fly by a database query that produces a quantile category to be used to symbolize. Place-tree is currently populated down to the country level.

5 Temporal controls

The timeline can be manipulated in three different ways. First, timeline sliders can be manually adjusted on either end, as illustrated below.



Multiple fixed time ranges are accessible by clicking the fixed time range link below the timeline. When first clicked, this option will set the time range to a default width of one month, as shown below.

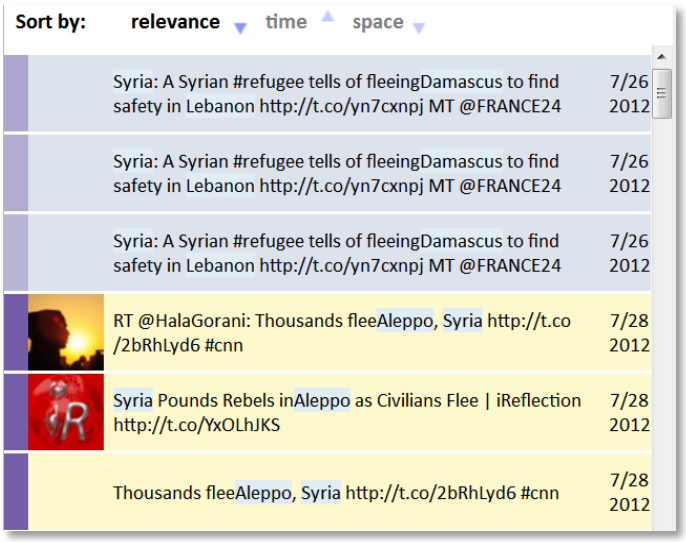
The width of the time range as well as its units ("one" and "month", respectively) are both adjustable. To adjust the number of months, hover over number 1 (a prompt saying drag will appear), click the mouse, hold the mouse button down, and drag. The timeline will adjust accordingly, as shown below.

You can also click on the label months to change it to week, day, or year. You can now either click on re-run time query link, or drag the time range to any part of the timeline to initiate the new query.

It is also possible to change the resolution of the timeline display. Use the radio buttons in the “Bin by:” control to the right of the timeline to set the width of the color bands to either one week (the default) or one day.

6 Tweet List

The tweet list (as shown in figure below) has two visual significations and three kinds of manipulation available to the user.



6.1 Visual significations

6.1.1 Tweet relevance

The narrow bar at the left of each tweet is color coded to indicate relevance to the query (with darker color depicting more relevant tweets), as estimated by the search engine.

6.1.2 Tweet locations

Locations that the system has identified in each tweet are now highlighted; this is currently used to help users visually recognize tweets relevant to places they are interested in but also to help users identify miscodings of places (see section 5.2.3 below).

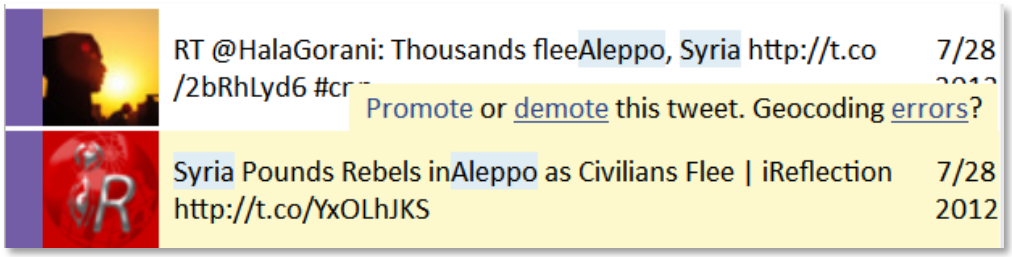
6.2 Manipulations

6.2.1 Sorting tweets

Tweets in the tweet list can be sorted by relevance rank, by timestamp, or by their location (currently labeled as “space”). The *location sort* is done based on distance of “about” locations contained within individual tweets from the current map center. So, at present, to sort tweets based on their proximity to the place of interest, it is necessary to center the map on that place first.

6.2.2 Promotion-demotion of individual tweets

Specific individual tweets can be temporarily promoted (moved to the top of the list) or demoted (moved to the bottom of the list). This is accomplished by first bringing the mouse over the specific tweet, at which point a line saying "Promote or demote this tweet. Geocoding errors?" will pop up, as shown in the figure below.

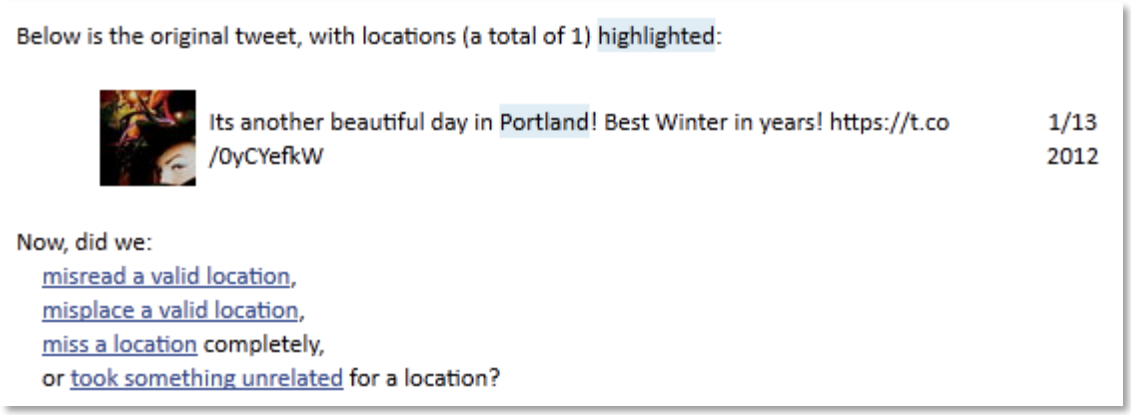


Clicking on either promote or demote will push this specific tweet to the top or bottom of the tweet list, respectively. The results of promotion and demotion will only be visible when the tweet list is sorted by relevance rank, and will be hidden when sorting by time or space.

6.2.3 User input on geocoding errors

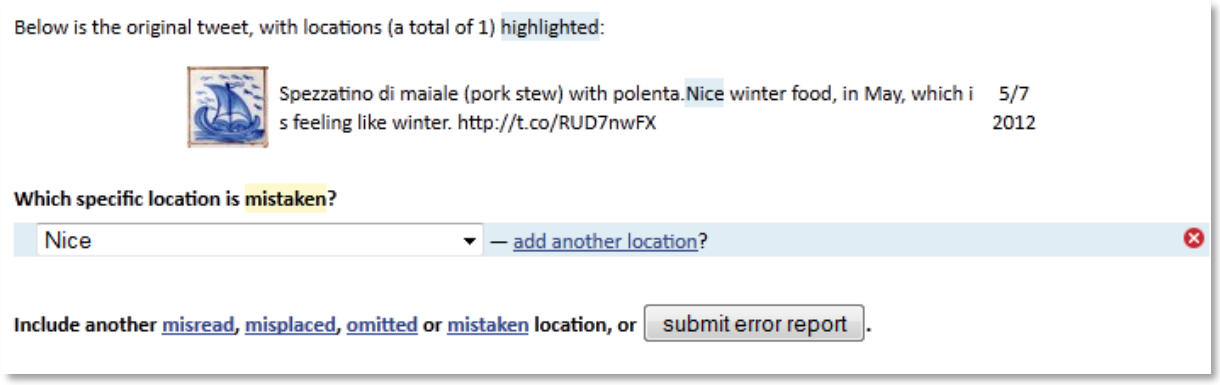
This feature is not completely implemented as of the date of this report. At present, the interface has been implemented and the raw feedback data is collected on the server, but the system does not yet process any of the data the user provides. This feature is described here as a preview of the upcoming functionality and because of its central role in the process of interactive geocoding, a significant area of future work. A detailed description of the geocoding error report functionality has been previously circulated as an independent document. For the sake of convenience, an abridged summary of this feature's functionality is presented below.

First, the user brings the mouse over the specific tweet, at which point a line saying "Promote or demote this tweet. Geocoding errors?" will pop up. User then clicks the on the "Geocoding errors?" link that will bring up a pop-up window that has the geocoding report controls, as shown in the figure below.

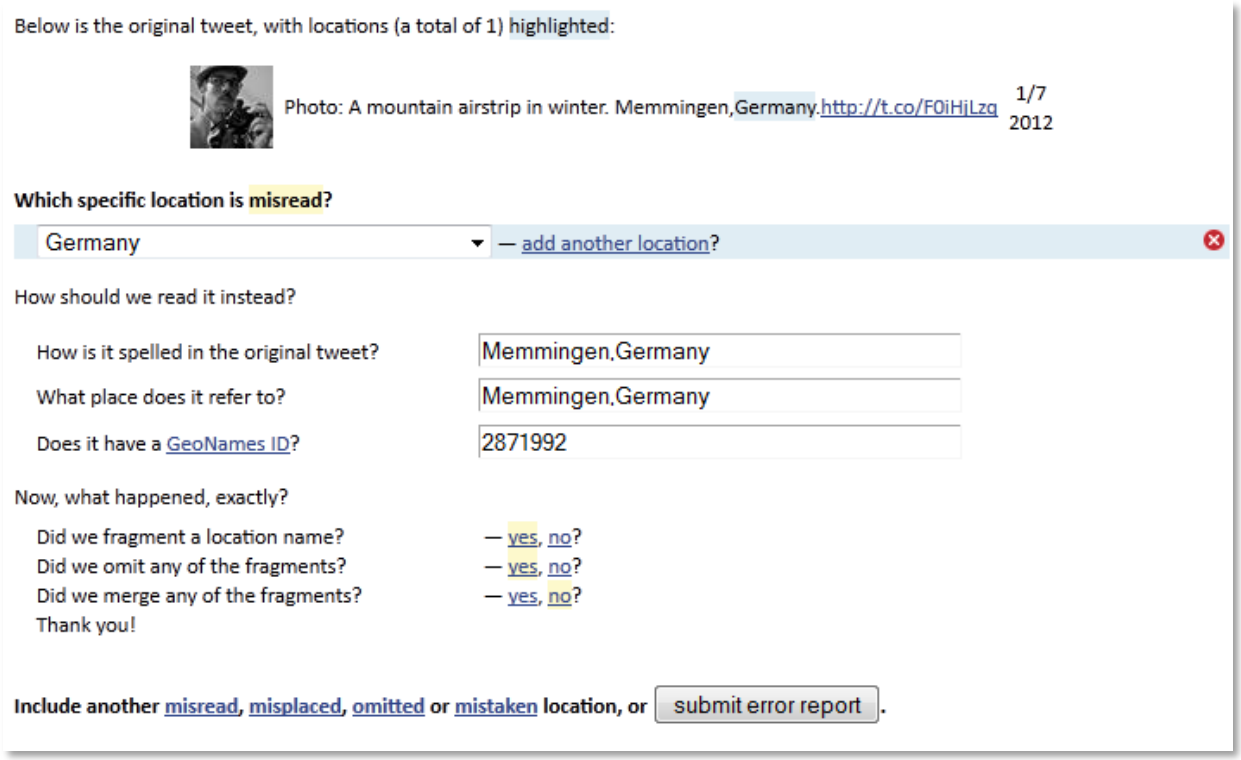


At this point, the user is prompted to select any type of error they would like to report, by clicking on one of the highlighted links. For example, "taking something unrelated" for a location implies that a regular word was taken

to be a valid place name. Once the user selects a particular type of error, the UI will be automatically expanded to incorporate user input, as shown below.



More complex error cases will request user input concerning the original spelling of the location, its *referent* (e.g. "Georgia" might refer to one of the US states or to a country of its own), as well as a questionnaire-based explanation of what exactly went wrong. A GeoNames ID lookup tool is also provided, activated by a click on the "GeoNames ID" link. An example of a more complex error report is presented below.



The geocoding error report can be expanded at will – all the user needs to do is click on the type of error they want to add ("misread", "misplaced", etc.) once more, using the links at the bottom of the report.

Once you start working with the geocoding error tool, a line saying "This report will be filed anonymously. Would you like to log in instead?" will appear at the bottom of the report. If you click the log in link, you will have a chance to pick a pseudonym that can later be used to link your reports together into a cohesive story. The pseudonym you pick is stored using the HTTP Cookies technology and can be changed at any time.

7 Map

The map includes several actions that enable users to focus attention on places and regions.

Mousing over a point on the map will highlight the tweet(s) associated with the point in the tweet list (if they are visible). Similarly mousing over a tweet in the list will highlight the point symbols associated with it. Tweets frequently refer to multiple locations at the same time. Thus, for each location, a list of co-occurring locations can be built. When a particular location is highlighted on a map, this list is retrieved and shown in the form of connecting lines between the original and the co-occurring locations, as shown in the figure below. The width of the line depicts quintiles of frequency for connections (bold lines represent more connections).

Clicking on any point symbol on the map will bring the tweet(s) associated with that point to the top of the list. Multiple tweets can be selected in succession – the tweets in the tweetlist will be highlighted using different colors for each group of tweets to help identify the original sequence of selections. Clicking the same point symbol again will deselect that particular tweet, while clicking in a blank space with turn all of the promotions off and put the list back to its default. When a place is clicked that has connections, the connections remain visible as long as the place is highlighted. Thus, it is possible to click on a few places in succession (without click a blank space to clear the selections) to build up a network of connections from a few selected places.

As discussed above, an *Alt+Click* combination, when performed anywhere on the map, will launch a new query using the current query terms and a spatial constraint that brings back the 1000 tweets closest to the point clicked. Keep in mind that the 1000 most relevant tweets returned might have mentions of locations outside the desired region.

8 Place-tree

The place-tree (as introduced in Section 4.3 above and shown in the figure below) has a number of user controlled features. Users can toggle between a place-tree that shows the full set of locations (as described by the GeoNames hierarchy) and a "pruned" hierarchy that only shows places that match the user query parameters. The switch is performed using the **"Switch hierarchy"** button.

Users can select one or more places of interest using check boxes positioned next to them, which would highlight the tweets related to that particular location in the tweet list. Use **"Drop selection"** button to clear all of the check boxes set.

Finally, although the capacity to launch queries based on GeoNames IDs of the features selected in the place-tree has been put in place, the server side of this functionality has only recently been implemented and needs further testing. This capacity will be added in the upcoming release. Thus, the **"Query selected"** button is currently disabled.

9 Tag Cloud

The tag cloud (as shown in the figure below) displays the list of extracted entities (e.g. place mentions) that are most frequently referred to in the query results. The size of the words in the tag cloud is proportionate to the number of mentions, and the words themselves can be clicked in order to filter the contents of the tweet list. Clicking again, de-selects those tweets.

10 Summary

Sections 1 through 9 above demonstrate the implementation of a number of methods that can be used to analyze place-relevant text information in tight integration with interactive maps. Some of the methods (such as the OpenLayers-based heatmap, point symbol display, and connecting lines display) overlay the relevant information on the map itself, whereas others (such as the place-tree hierarchy component, tweetlist and tag cloud component) are an example of a coordinated view approach due to their strong coupling between each other and the map display. Finally, a number of tools (such as promote/demote and geocoding error report functionality) fall into the category of iterative, user-driven visualization techniques.

Despite the overall improvement in the quality of visualization and analysis tools, as well as the introduction of a number of new components, much of the current progress should be seen as setting ground for further research

and design iterations. As described above, current implementation of SensePlace2 is primarily query-driven, and provides detailed exploration of the top 1000 tweets corresponding to the given query as a proof of concept. However, a more thorough implementation of the Information Seeking Mantra mentioned above would call for longitudinal exploration capacity and multi-tier drill-down beyond what was possible to implement as part of the work planned for Component 2 (e.g. picking out a user from the tweetlist, exploring all of their tweets, following a particular one through its retweets, etc.).

In addition to the development of novel visualization techniques, much work needs to be done on the server side of the system, including the information processing pipeline and the related middleware. With improved flexibility on the server side, further refactoring on the UI side of the project will also be necessary to fully utilize the potential of the data available.

Finally, a user study that is performed in Component 3 of this project is expected to provide broad insight on the usability of the SensePlace2 UI and on strategies to increase usability and utility of the system.

References

Cockburn, A., Karlson, A., & Bederson, B. B. (2008). A review of overview+ detail, zooming, and focus+ context interfaces. *ACM Computing Surveys (CSUR)*, *41*(1), 2.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. *Visual Languages, 1996. Proceedings., IEEE Symposium on* (pp. 336–343).