NewsViews: An Automated Pipeline for Creating Custom Geovisualizations for News

Tong Gao, Jessica Hullman, Eytan Adar
School of Information
University of Michigan
{gaotong, jhullman, eadar}@umich.edu

Brent Hecht
Computer Science & Engineering
University of Minnesota
bhecht@cs.umn.edu

Nicholas Diakopoulos
School of Journalism
Columbia University
nad2141@columbia.edu

ABSTRACT

Interactive visualizations add rich, data-based context to online news articles. Geographic maps are currently the most prevalent form of these visualizations. Unfortunately, designers capable of producing high-quality, customized geovisualizations are scarce. We present *NewsViews*, a novel automated news visualization system that generates interactive, annotated maps without requiring professional designers. NewsViews' maps support trend identification and data comparisons relevant to a given news article. The NewsViews system leverages text mining to identify key concepts and locations discussed in articles (as well as potential annotations), an extensive repository of "found" databases, and techniques adapted from cartography to identify and create visually "interesting" thematic maps. In this work, we develop and evaluate key criteria in automatic, annotated, map generation and experimentally validate the key features for successful representations (e.g., relevance to context, variable selection, "interestingness" of representation and annotation quality).

Author Keywords

Narrative information visualization; interactive maps; online news; geovisualization; text summarization.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Online news covers a gamut of information on economic, social, and political topics (among others), spanning local, national, and global scales. Its digital nature affords a range of novel mechanisms to provide context, navigation, and dissemination of additional information. Media such as images, videos, and information graphics can be created and integrated into news article presentations to augment the reading experience. Of particular interest are the sophisticated data visualizations used to present relevant data and summarize additional context for these articles. By providing static and interactive data "views," visualizations offer a more tangible depiction of trends explicitly mentioned or indirectly implied in an article. For example, a thematic map supplementing an article about geographic variation of economic mobility in the United States might show the geographic trends at a level of detail that is difficult to obtain only through text (Figure. 1). A news reader can query the map based on the article content and data, comparing values for referenced locations with particularly high or low income mobility, or making personally-relevant queries concerning income mobility in her hometown. These visualizations are often annotated [13, 25] to guide the reader to key points and add context for the story being told.

The annotated visualizations that appear in news contexts are typically the work of skilled designers and data journalists. Among other tasks, the designer must find relevant data that enhances and complements the article, select appropriate representations and encodings, maintain various design possibilities and select useful and appropriate annotations. Each decision in this pipeline has an effect on sub-

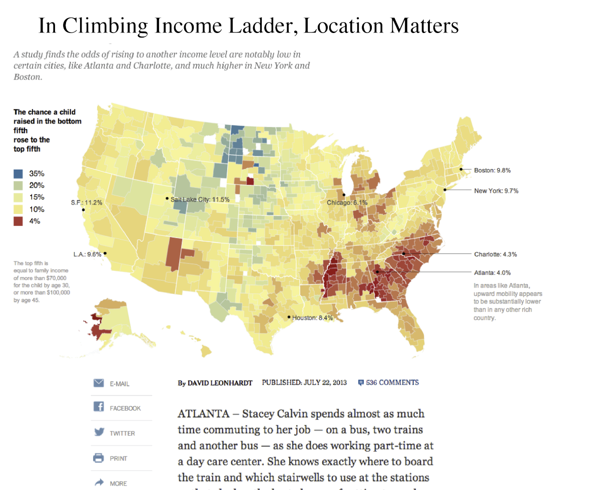


Figure 1: Thematic map by the New York Times describing geographical differences in the likelihood of moving up an income bracket [19].

CHI 2014, April 26–May 1, 2014, Toronto, Ontario, Canada.
Copyright ©ACM 978-1-4503-2473-1/14/04...$15.00.
http://dx.doi.org/10.1145/2556288.2557228

sequent steps and frequently requires time-consuming iteration. As such, resource limitations and process complexity constrain the creation processes of professionals, making it challenging to provide visualizations for the large collections of articles available on news websites today.

Automation has been employed in news visualizations, yet these few solutions tend to be specific to small subsets of news articles. For example, applications generate annotated stock time series [12] or graphics to accompany articles that follow a strict narrative template [22].

We explore how a complex visualization *pipeline*, or sequence of decisions and algorithms for solving them, can be automated to scale the creation of popular types of news geovisualizations. Our work contributes to automated news visualization research by identifying key criteria for a news map generation pipeline. We describe the *NewsViews* system, which implements this pipeline to automatically generate visualizations to accompany news articles using geographic information from multiple sources. Criteria include pairing relevant data variables to article text, finding locations of interest, choosing relevant annotations, setting the map extent, and modeling maps' visual interestingness.

We secondly contribute a demonstration of how these criteria can be operationalized, employing text mining and extraction techniques as well as visual-spatial pattern analysis to realize the NewsViews pipeline. NewsViews identifies relevant features of the article (topics, locations, etc.) automatically, and uses these features to select from hundreds of data variables and ~670,000 spatiotemporal "cases" (i.e., database cells). Multiple possible "views" (each representing a set of salient columns from a multi-topic table database) are generated and visualized. A novel ranking technique selects the visualization to maximize the relevance of selected data and annotations to the article and leverage statistical features to align with users' perceptions of map "interestingness." Our final contribution is an evaluation of users' reactions to our operationalized criteria. We evaluate our criteria as operationalized both independently and in combination, including a novel perceptual evaluation of Moran's I as a measure of visual "interestingness."

RELATED WORK

Narrative Visualization

Research around narrative visualizations—visualizations that guide users' interpretations using forms of visual, textual and procedural rhetoric [13, 25]—informs the design of NewsViews. Many exemplary artifacts motivating scholarly discussion come from news settings, where interactive visualizations presented with text articles help explain news [5]. Text annotations play an important role in guiding users' interactions and interpretations [25, 13]. Annotations are used in two primary ways: to emphasize data-based observations ("observational" use), such as a maximum value in a series, or, as applied in NewsViews, to add information that is not otherwise presented ("additive") [12].

Typically, the annotated narrative visualizations that appear in online news contexts are the work of designers with expertise in graphics, statistics, and journalism. The time these experts spend creating graphics also makes it difficult to scale their process to the massive amounts of new and archived news content, motivating automated solutions.

Automatic Visualization Generation and Annotation

Various recent projects contribute methods for automatically generating or supplementing visualizations, such as by depicting a chain of news stories connecting two articles [27], or summarizing relations between scientific papers using metro-style maps [26]. Google News Timeline and Google Finance present automatically-generated visual summaries of articles, though typically using relatively simple chronological parameters. Narrative Science (http://narrativescience.com) produces narrative style reports and graphics to walk end-users through datasets of interest. Visualization researchers have contributed methods for adding annotations to existing visualizations, including graphical and textual annotations to emphasize trends, maximum points, or means, among others [16], and text annotations on interesting trends in point data (e.g., [15]).

Our earlier system, Contextifier [12], produces narrative visualizations for online business news. Textual features of a news article about a company are used to generate a customized line graph of that company's stock. Additive annotations drawn from *other* articles provide information relevant to the company and context, yet external to the original article or data. These annotations are selected and placed using a combination of topical relevancy features, features suggesting important (historical) company events, and saliency features computed on the data series. Most notably, whereas Contextifier focused on one data type and one specific instantiation (i.e., a stock time series) NewsViews is designed to work with articles from arbitrary domains, selecting appropriate data of various types (i.e., georeferenced data, time series) and between possible visualizations. Like Contextifier, NewsViews generates annotation content from a news corpus but can also support observational annotation of extreme values (e.g., min/max).

Wu et al. [21] proposed the semi-automated MuckRaker system for connecting news readers to a database of relevant context using a visualization interface. Unlike NewsViews, MuckRaker leverages the crowd to improve automatic presentation features, yet still calls for explicit input from the user. Other work focuses on the generation and searching of very large tables of structured data, but does not concern visualization [4].

Geovisualization and News

Maps often appear with news articles to improve readers' understanding of a story's geographic context, to make geographic trends personally relatable, or simply to attract readers' attention [5, 9, 31]. A survey identifies maps as the most prevalent visual format used in a recent sample of narrative news contexts [12]. Presenting maps with articles

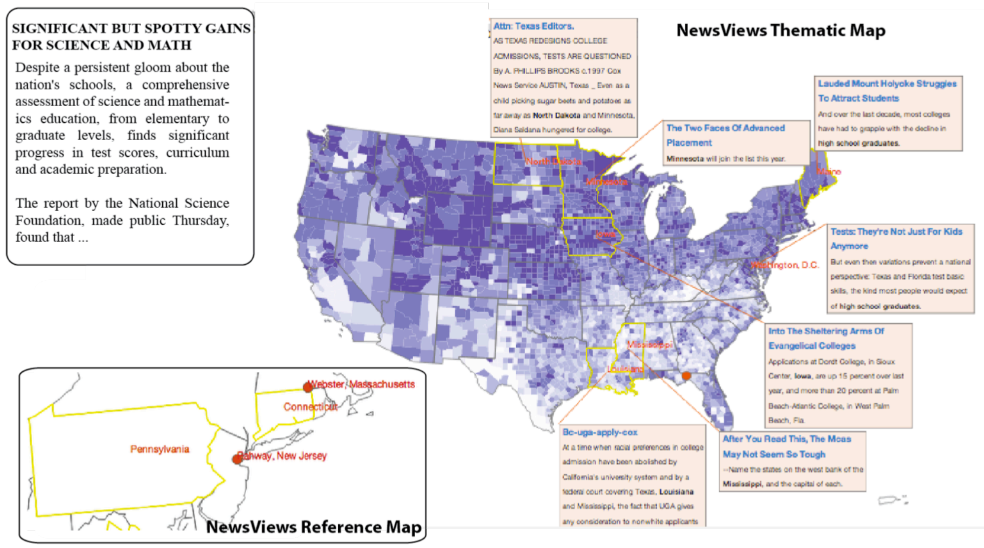


Figure 2: The NewsView interface, depicting an input article and custom annotated thematic data. An example of a reference map is shown to the lower left.

increases a news reader's geography knowledge via dual-coded learning [9]. NewsViews automatically scales these benefits to large numbers of articles.

The cartography literature divides maps into two broad categories: *reference*/*locator* maps and *thematic* maps [29]. Both forms frequently appear in news geovisualizations [5]. Reference maps are dedicated to the communication of specific locations such as boundaries or points of interest. Thematic maps are used to "emphasize the spatial pattern of one or more geographic variables" [29]. NewsViews supports both types of maps and intelligently chooses the appropriate type given a news story.

SYSTEM OVERVIEW

We briefly describe the user experience and high level architecture below.

User Experience

A news reader interacts with NewsViews as she reads a text article as part of her normal interaction with a news website. If the article mentions one or more locations, a customized NewsViews visualization is presented to supplement her reading alongside the article. If the article topic is best contextualized with data, such as in the case of an article about the results of a nationwide evaluation of educational performance, the presented map is a customized thematic choropleth map, showing education attainment by county (see Figure 2). In this view, the thematic map depicts the education attainment of residents by county by coloring areas in proportion to the measured level of the variable during the time frame implied by the article text. In other cases, the article may not imply relevant data (i.e., thematic layer), but may still mention locations that are likely to be unfamiliar to the user. In this case, the presented visualization is a "reference map" which the user can leverage to understand the spatial context of mentioned locations. For example, an article about a marathon in Nixa, Missouri may yield a zoomed in map of Missouri that places Nixa in the context of better-known nearby cities.

The map supports analysis in several ways: (1) details-on-demand [28] allow the user to compare the specific levels of locations mentioned in the article, providing her a more detailed perspective than can typically be gained from the article alone, (2) she can infer the broader spatial distribution of educational attainment by examining levels across the entire U.S., and (3) she can zoom to specific areas of personal interest, like her hometown or current residence.

For news readers to fully understand an issue often calls for more textual context than is available in a single article. Customized annotations can provide this additional context. For example, the user may have little prior knowledge on educational trends in general or the specific tests mentioned in the article. Or, she may be curious for more information describing why some regions or states have much lower scores. Annotations affixed to map locations provide additional information such as location-specific explanations of data, or details on particular tests used. The titles of the articles referred to in annotations can be used to navigate to read these articles in full, and to view a new map customized to that article instead. Thus, the reader can navigate to related articles via the graphical interface as an alternative to using text-based lists or search features.

Other features. Observational annotations that simply point to extreme values are also common in narrative visualizations in the news [12]. NewsViews presents the reader with annotations of extreme values in cases for maps where several "outlier" locations display uncommonly high or low values (see Fig. 4, bottom). We discuss extensions of the system to create further visualizations below.

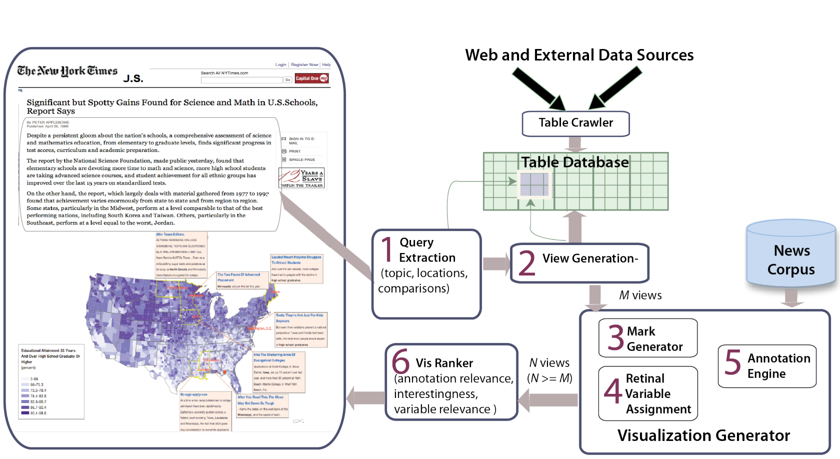


Figure 3: The NewsViews system architecture.

NewsViews: Visualization Pipeline	
1	What are the topic, locations, and implied comparisons?
2	What table views (data graphics or reference maps) match extracted query?
3	What mark types (graphical formats) are compatible with each of the M views?
4	What is the optimal retinal variable assignment given the data in each of M views?
5	What annotations add context to the topic and visualizations?
6	Which annotated visualization maximized relevance and visual interestingness?

Table 1: NewsViews visualization pipeline.

The NewsViews Corpus

NewsViews' ability to produce high quality visualizations for news is driven by several large databases of textual and tabular information, which first describe here.

News Corpus

The news corpus we use consists of a set of ~1.9 million news articles from the New York Times ranging from July 1994 to December 2006, covering all sections of the newspaper [24]. Since we focus on location-related context, we first matched articles against a complete list of U.S. counties, states, cities, and region names. This left ~1.5M articles that mentioned at least one U.S. location. We subsampled ~1M of these for the NewsViews corpus.

Table Database and Table Crawler

The table database is a collection of temporally and georeferenced variables, such as state and county-level unemployment, disease statistics, or statistics describing the prevalence of different industries. Our system supports "found" tables (e.g., Excel files); however, these are often noisy and uncontrolled. For evaluation purposes we used two primary sources of data: the United States Census and Freebase. We extracted tables from Freebase by "pivoting" this database. The Freebase structure includes a high level entity as the main page (e.g., San Francisco, CA) with RDF triplets that indicate a relationship to the higher-level entity (e.g., "pollution in") and value (the number, units, and a measurement date). We extracted these triplets for each US location and generated one table for each variable (containing all locations for which we had a measure).

We manually augmented this collection with recent tables from the Statistical Abstract of the United States (SAUS) [32] and the Bureau of Labor Statistics. Because keys for locations change across data sources (e.g., San Francisco can be referenced using a name in one table, a FIPS code in another, and a GeoName id in a third), we built a mapping between these IDs by combining the Freebase, Wikipedia/DBPedia, and the GeoNames dataset.

From these datasets we selected 155 "high quality" tables vetted by one author of this paper (with expertise in geography and cartography). This also allowed us to identify and fix variables that were best expressed as normalized quantities (e.g., percent unemployment rather than raw unemployment counts). We ensured that variable names were "cleaned" as these are used by our algorithms (e.g., unemp. would be transformed to unemployment). The resulting tables spanned topics including education, employment, income, population, transportation, agriculture, climate, manufacturing, diseases and health, wealth distribution, and science and technology.

The NewsViews Pipeline

The NewsViews pipeline negotiates six main decision points (see Fig. 3, Table 1) that leverage the corpus and table database. We focus our discussion of the pipeline here on the challenges in creating thematic maps to accompany articles that contain *toponyms*, or place names, as well as geographically anchored entities (e.g., *Harvard University*). We note that dependencies on location and variable selection impact "downstream" decisions such as annotation relevance and visual interestingness. We describe below how and why NewsViews prioritizes these decisions.

1. Query Extraction

The query extractor identifies important attributes that are subsequently used to infer related variables (e.g., unemployment by county) and entities (e.g., locations). We follow an approach based on the "inverted pyramid" of journalism, which states that the first several sentences of a news article often contain the main point of the article ([7], applied in [12]). Noun phrases from the first three sentences are extracted as seeds for topic identification.

To identify locations mentioned in an article, the query extractor extracts U.S. state names, county names, city names and organization names. We employ the Wikifier system [18], which identifies all entities in text with links to their Wikipedia page. This is beneficial because most Wikipedia pages for entities with geographic coordinates record these in the page. The query extractor thus "anchors" detected entities to a geographic footprint.

Wikifier outputs multiple possible links for ambiguous entities. To boost the accuracy of NewsView's location tagging accuracy for these cases, we combine Wikifier's output with that of OpenCalais [23] (a second named entity detector). OpenCalais is used to filter entities tagged as locations by Wikifier but persons by OpenCalais.

2. View Generation (Column and Row Selection)

Key to a successful visualization is the choice of an appropriate data variable. The view generator takes the information outputted by the query extractor (consisting of 'topical' noun phrases, identified locations, plus the article time stamp) and applies further analyses to generate "data views" (representing a selection of columns and rows) on a large table database. The NewsViews database consists of columns that reflect a particular piece of data (e.g., unemployment, education, etc.). Rows are keyed by location for US states and counties (with counties as a default). If there are multiple samples of a variable from different periods (e.g., unemployment from 2009, 2010, 2011, etc.) a different column is assigned to each. Retaining historical values is useful as it allows us to generate other visualization types (e.g., a time series for a given location) but also allows us to default to data from the time the article was written—a 2005 article on unemployment plot data from 2013.

A set (of size M) relevant views identified by the view generation step is then be passed to the visualization generator. The specific mechanisms for using the query information to select from all possible variables are detailed below.

3, 4, and 5. Visualization Generation

The visualization generator takes the filtered set of M views identified by the view generator and uses visualization best practices to generate a set of N annotated visualizations with D3. A *mark generator* (3) first assigns the best visualization mark type (e.g. point, line, polygon). In the default NewsViews implementation, these are polygons depicted using a Mercator projection. However, the pipeline is extensible to other mark types such as lines, bars, or small multiples, discussed further below. A *retinal variable assignor* (4) maps the selected data to a retinal variable (e.g., sequential color scheme with optimal binning; see Thematic Map Generation for details). The retinal variable assignor also identifies the default geographic extent (zoom level). An *annotation engine* (5) selects the descriptive annotations for (explicitly mentioned) map points from our news corpus. The generator may produce more than M visualizations (i.e., $N \geq M$) to allow for cases where more than one visualization of a table view is possible (e.g., when the annotation content differs between two maps, different bins are used, or different zoom levels are proposed).

6. Visualization Ranking

The *visualization ranker* ranks the N visualizations produced by the visualization generator. Three criteria are used: the quality of the variable selection as captured by pointwise mutual information (PMI) between article text and variable labels, the annotation relevancy as captured by cosine similarity, and the visual interestingness as captured by Moran's I. Our ranking (detailed below) prioritizes variable match to article (via PMI), followed by visual interestingness. Our evaluation of annotation relevance and visual interestingness (Fig. 7) supports this ordering.

ALGORITHMS FOR FINDING THE RIGHT MAP

Below, we detail the specific algorithms used to model each step in the NewsViews pipeline for map generation (Fig. 3).

Identifying the Best Variable for a Thematic Map

A critical decision in the map creation pipeline concerns the data to be depicted with an article. The noun phrases from the first three sentences of an article (e.g., "job hunters" or "undergraduate degree") rarely map directly to the name of variable likely to be most suitable (e.g., "percent with bachelors"). We devise an algorithm to negotiate this decision using a measure of mutual information.

We rank the 155 variables (i.e., column descriptors) by their relevancy, calculating the co-occurrence score between the noun phrase and variable name weighted by Pointwise Mutual Information (PMI) [2]. PMI measures how often related two terms are by comparing how often both terms appear together in an article together versus independently. The more articles that mention both phrases simultaneously (i.e., the variable name and the extracted noun phrase), the higher the PMI. In practice, we convert the original variable description into a sub-phrase set and determine the mean PMI between these and the article-derived noun phrases.

Isolating descriptive variable names: Each variable is represented as a descriptive phrase, e.g. "fast-food-restaurants-by-county," which may be too specific to match any given article. To improve recall we automatically identify the noun phrases from the original variable string, e.g. 'fast food restaurants', 'fast food', 'food', 'restaurant' and 'county'. For each phrase, we count the matching documents in our corpus (as a fraction of corpus size). We discarded terms with relatively high frequencies (>0.05) such as 'county' (0.057), and those with relatively low frequencies (<0.0001) such as 'educational attainment' (0.00009). We retained those with moderate score: 'fast food restaurant' (0.00052) in the variable 'fast-food-restaurants-by-county', and 'bachelor's degree' (0.00054) in the variable 'educational-attainment-25-years-and-over-bachelor's-degree-or-higher.' The filtered set of noun phrases is retained (we refer to this set as Variable Phrases, or *VarP*).

Capture Potential Topic From Articles: For a given article, all the noun phrases from the first three sentences are extracted from the query extractor. We remove phrases that have low *tf-idf* scores (<0.001), as these are less likely to carry important information about the article topic (weighting is calculated based on global frequency of the phrases in the news corpus collection). We refer to the remaining noun phrases as "NP terms."

Calculate PMI: We can then calculate the PMI between each NP term and VarP term and aggregate these as the mean PMI for all pairings. For efficiency, all articles in our dataset are indexed in a Lucene database. For each pair, we issue 3 queries, one for NP, one for VarP, and one for both VarP and NP simultaneously (the number of matching articles are retained as NP_{count}, $VarP_{count}$, and NP_VarP_{count} respectively). From this, we calculate PMI as follows:

$$PMI(NP, VarP) = log_2 \frac{P(NP, VarP)}{P(NP)P(VarP)}$$

$$= log_2 \frac{NP\_VarP_{count}/N}{(NP_{count}/N)\times(VarP_{count}/N)}$$

The mean PMI is calculated by determining the PMI for each phrase pairing (i.e., one item from NP and one from VarP) and normalizing by the number of such pairings (i.e., NP_{count} * $VarP_{count}$. Variables are ranked by their mean PMI relative to the extracted noun phrases. The intuition here is that the higher a variable's PMI value is relative to the article text, the more relevant the variable to that content.

Thresholding to Focus Variable Relevance: Given our expectations that the match between article content and a relevant data set will most drive map usefulness, we threshold the PMI-ranked list. This involves finding the threshold based on where PMI values decrease sharply in the ranked list and then filtering to the high PMI articles.

Reference Map Default: NewsViews supports reference as well as thematic maps. A heuristic is used to create reference maps if no PMI score for any variable surpasses a threshold (set to 2.5 after qualitative experimentation). If the threshold is not met, a reference map is created. The map is zoomed to the locations mentioned in the article, annotated with simple place-markers (Fig. 2, lower left).

Generating Thematic Maps

Having selected a set of georeferenced data from the table database, a retinal variable assignor determines the optimal mappings of a retinal variables (e.g., color, size, etc.) to visualize the data on a map. In the current implementation, we classify the data into 7 stepwise classes with the Jenks natural breaks classification method [14], an iterative data classification method to determine the best binning of values given the observed distances between sets of values. Following Brewer's color guidelines for mapping and visualization [3], NewsViews selects sequential schemes for the Jenk's derived classes in which lightness steps dominate, using light colors to depict low data values and dark colors for higher values. Future work might incorporate alternative schemes (e.g., gradients or semantically-driven colors [20]).

Identifying Related Locations and Geographic Extent

Primary Location Identification: Defining an appropriate map extent (or default zoom region) allows users to focus on the critical area that is relevant to the article. We first identify the "primary location" that represents the main location of interest for that article by extracting "seed locations" from the title and the first three sentences. The occurrence of each seed location within the article is counted. Seed locations are ranked based on occurrence, and the most often mentioned seed is assumed as primary.

Identifying Related Locations: We then detected the relationship between the primary and other locations (l_i, the i^{th} location in the set of other locations mentioned in the article text). Three distance measures are calculated and combined. A spatial distance captures the Euclidean (coordinate) distance between location l_i and the primary location. A hierarchy distance captures the distance between the two locations on the geo-hierarchy tree, prioritizing local extent. This allows us to capture the idea of "sibling" relationships (two counties in Virginia), enclosure relationships (Michigan → Washtenaw County → Ann Arbor), and other longer range relationships (e.g., Ann Arbor, MI and Alexandria County, VA). Different weights can be assigned to different "hops," (e.g., state to county versus county to city) though we use an equal weight. Finally, a weighted location co-occurrence distance corresponds to the number of times the two locations are mentioned together in the same corpus. We use the $PMI(l_i, l_{main})$ function, as described above, to calculate this weight.

The three measures are combined in a final location similarity score for each location and the primary location $sim(l_i, l_{primary})$ by dividing the weighted location co-occurrence difference (which reflects a strong relationship between the two locations), by the sum of the spatial and semantic distances (where larger numbers indicate dissimilarity). Other implementations of these three forms of location similarity are possible. We stress the importance of the gross relationships between them over specific operationalizations and combination functions.

Filtering Locations and Setting Extent: We experimented with different thresholds for filtering the location set. Specifically, we kept the locations that had high scores ($sim(l_i, l_{primary}) \geq 0.1$), and discarded those with low scores ($sim(l_i, l_{primary}) < 0.1$). We set the map extent to the area that exactly contains all the high scoring locations (as well as the primary location).

Selecting and Placing Text Annotations

NewsViews supplements each generated thematic map with *additive* annotations—textual descriptions that provide related information (context) about the map, data, and input article topic where the source of that information is not the data or input article itself [12]. That is, we find text in *other* articles that are related to locations and data on the map.

For example, NewsViews considers the related locations (calculated above) for which we have articles that focus on the location and variable of interest. Thus, if the context article is about unemployment in Springfield, MA, NewsViews will identify Westfield, MA (a nearby town that often co-occurs in articles about Springfield) and Providence, RI (a similarly sized New England town that also co-occurs in Springfield-related articles). NewsViews will then isolate news articles about these locations (Westfield and Providence) that focus on unemployment (preferring those that also mention Springfield). From this set of related articles, content will be extracted as annotation material.

Finding Location-Topic Article Matches: For each related location we query the corpus for articles that mention the location *and* the topic words (recall, the "VarP" term). These articles are sorted based on their cosine similarity (on the bag-of-words article representation) with the input article, and the article with the highest similarity selected. These search steps ensure the related articles are in fact related to the variable of interest and the context.

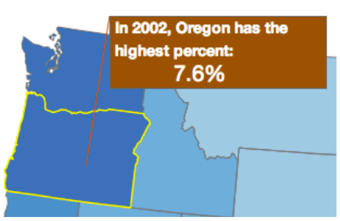
Descriptive Sentence Selection: Once the best article has been identified for a location, a descriptive sentence is selected from the article text by extracting all sentences that mention the location and ranking the sentences by topical relevancy. To do so, we identify the noun-phrase terms in the possible sentence and calculate their mean PMI given the VarP term. The sentence with the highest mean PMI is used as annotation content. This further enforces topic-specificity by not only selecting articles related to the topic and location but finding the best sentence from that article.

Annotation Placement: Many possible algorithms exist for placing map annotations without overlap. We devise a simple custom placement algorithm to match annotation realizations (including the parameters of the text box) to available screen space.

Annotations need not be driven by content extracted from other articles. For thematic maps with a small number (<2) of relevant annotations, NewsViews also creates *observational* annotations, such as pointers to the highest and lowest points with corresponding values (Fig. 4, left).

Visual Interestingness (Saliency) Analysis

One of the benefits of thematic cartography is its ability to communicate regional patterns better than other visualization approaches [29]. We hypothesized that a metric that captures the extent to which a spatial distribution exhibits regionalization would serve as a good proxy of that map's interestingness. Measures of spatial autocorrelation accomplish this. Broadly speaking, spatial autocorrelation refers to the extent to which two spatial features (e.g. counties) nearby each other are more similar than those that are farther apart. We compute a measure of spatial autocorrelation called Moran's I for each of our 155 variables. Given that little is known about the relationship a given spatial variable's properties and the "interestingness" of a map of that distribution [10], we evaluate how Moran's I affects user perceptions in Evaluation, below.



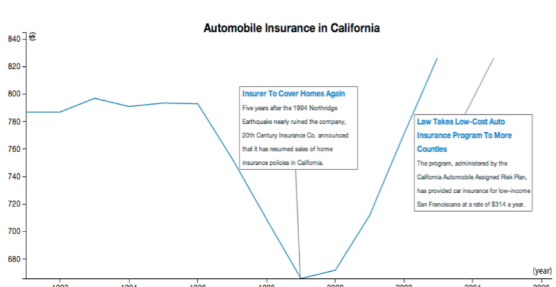


Figure 4: Observational annotation pointing to extreme value (left), and line graph generated for an article implying temporal comparisons (below).

Final Map Selection

The final map selection process seeks to find the map that is most likely to display highly relevant and visually interesting data with topically appropriate annotations to explain locations on the map. We threshold possible visualizations to include only high PMI scoring variables (e.g.., top 3). The map ranker computes Moran's I to predict the visual interestingness of each high PMI map, retaining this score. We then sort the possible visualizations by the mean annotation relevancy of the visualization's annotation set. The ranker selects as the final visualization that high variable relevance visualization that maximizes visual interestingness, while staying above a threshold annotation relevance.

EVALUATION

We expect at least four aspects of the algorithm to affect the perceived quality of NewsViews' maps: (1) precision and recall of article location tagging, (2) ranking and assignment of data variables, (3) selection of annotation content , and (4) prioritization of visually interesting maps.

The first two factors play a fundamental role in NewsViews' map selection process as they comprise the view generator. We evaluate two questions in separate Amazon's Mechanical Turk studies with human raters:

1) Do extracted locations align with human raters' assessments of important locations in articles?
2) Do the rankings produced in data assignment (column selection) align with the relative ratings of human raters evaluating the data to article match?

To investigate whether user perceptions' of map quality, interestingness, and relevance align with NewsViews methods for selecting annotations and interesting maps, we use a more "holistic" evaluation. In our study, participants evaluate annotation content, visual interestingness, and the overall quality of maps, which we generate using the combined NewsViews algorithms or a variant of the system in which a given feature type is not included. This allows for possible interdependence between the annotation relevancy and visual interestingness as contributors to the perceived quality of NewsViews maps (e.g., a rating of whether a map is visually interesting might be unintentionally influenced by a user's annotation rating).

Location Tagging Evaluation

We evaluate the precision and recall of NewsViews location tagging as implemented in the query extractor.

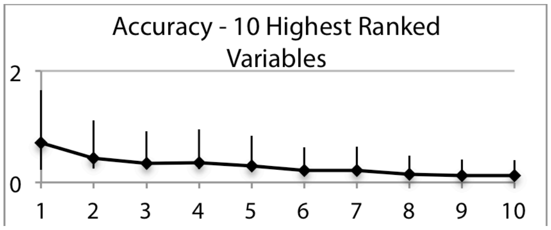


Figure 5: Accuracy of variable assignments to article for first ten variables ranked by PMI, aggregated across 66 articles. Error bars display standard deviation.

Task and Materials

We randomly selected 47 articles from the corpus and used the query extractor to tag relevant locations in each. One to ten locations were identified in each article. Four of the authors used assessed the precision and recall of each tagged location set (each article-location sets evaluated by at least 2 coders). Coders read the article, identified correctly extracted locations, and added up to 10 locations that were mentioned in the article but not extracted.

Results

Precision (the percentage of system-tagged locations that the raters confirmed as accurate) was 92.7% (σ=14.2) over the set of 47 articles. Recall (the percentage of locations that raters believed were of interest to the article but which were not captured by tagging) was 42.6% (σ=23.9).

Variable Assignment Evaluation

Task and Materials

We combined the 47 above articles with 19 more articles for which tagged locations were evaluated, and used the view generator to assign and rank data variables (table columns) for each of the resulting 66 articles. This created a list of 155 ranked variables for each article for which the PMI was know. A first group of Mechanical Turk workers read the articles in separate HITs, each of which displayed the article along with the top ten ranked variables using PMI. For each article, workers indicate whether they agreed with the classification of the variable as relevant to article content (yes/no question).

A second set of HITs based on the same 66 articles also displayed each with a list of variables. However, this time only six variables were shown, and workers were asked to rate the relevance of each variable to the article content using a Likert-style radio button list from 0 (Not Relevant at All) to 3 (Highly Relevant). The presented variable list included a mix of PMI-based ranks. Specifically, one variable was randomly drawn from the first five slots in the sorted list, a second variable from the first ten slots, a third and fourth variable from between slot 11 and the list midpoint, and a fifth and sixth variable from the lower half of the list.

In each case, U.S. workers with an approval rating of 95% and above were eligible to complete up to all 66 HITs for a base reward of $0.10. A worker received a bonus of $0.02

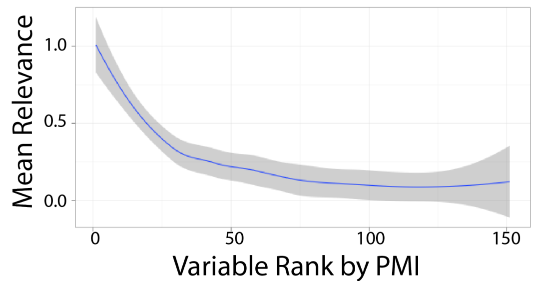


Figure 6: Loess curve applied to mean relevance by PMI rank for all variables, aggregated across 66 articles.

for each variable for which her rating matched that of the majority of other workers for that article and variable.

Results

30 and 41 unique workers took part in the two tasks, doing between 1 and 66 HITs apiece (μ =24) to provide a total of 6600 and 5940 individual variable ratings, respectively. Figure 5 depicts the aggregated accuracy for the relevance of the variables in the top ten positions from the first task. Results indicate a consistent negative trend between accuracy and rank by PMI.

We used the results of the second task's Likert relevancy ratings to examine how the perceived relevance of a variable decreases as the variables rank by PMI decreases for all possible variables. As shown in Figure 6 perceived relevance drops off at approximately rank 30.

Annotations, Interestingness, and Quality Evaluation

Task and Materials

We randomly selected 10 articles for location and variable match evaluation. For each of these articles, we created three map visualizations for comparison. A "full solution" (FULL) applies annotation relevancy and map saliency considerations to map creation. Possible variables are filtered to the top three matched variables by PMI to ensure that the maps present relevant data. We then select the variable out of these top three for which visual interestingness is likely to be predicted to be highest by using Moran's *I*. We compute annotations for each location by ranking all candidate articles matching a location and VarP term by cosine similarity and using the top article to create the annotation as described above.

Two partial solutions are also created to examine the relative contributions of having annotation content and prioritizing visually interesting maps. A "no cosine similarity" (NoCOS) map is generated for each article where article cosine similarity is not considered. Instead, a random article is chosen from the unranked list of candidate articles and used to create annotation content. A "no saliency" solution (NoSAL) also follows the same process as the FULL version, but with one exception designed to allow us to test very low saliency maps. We held all aspects of the FULL solution constant, including the labeling of the data variable, but presented randomly-generated artificial data in

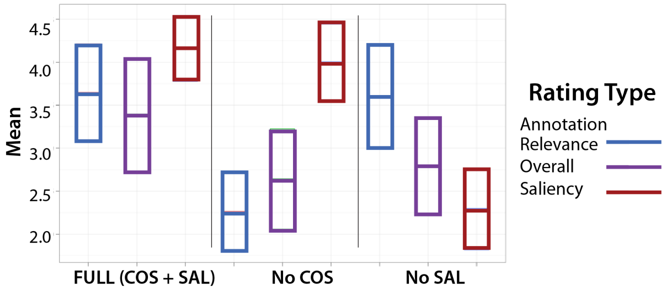


Figure 7: Mean Ratings by Feature Types. Crossbars depict standard deviation.

place of the true values in the map but labeling the data with the name of the actual data variable displayed in the FULL version. The intention of this was to see how much the spatial autocorrelation that Moran's *I* detects affects users appraisals of how visually interesting a map is.

Each of 10 users completed a rating task that presented the 30 article-visualization pairs in random order. Users were made aware of the possibility that the map data was artificial. The task asked the user to read the article and examine the map, answering three questions:

1) *Annotation Relevance*: How relevant is the content of the annotations to the article topic and details?
2) *Visual Saliency*: How visually interesting is the map? (Do the visual patterns incite your curiosity?)
3) *Overall Rating*: How useful is the map overall in helping explain the content in the article?

Results

10 university graduate students earned $15 for voluntarily participating via an online interface. Figure 5 depicts means for each of the Annotation Relevance, Visual Saliency, and Overall ratings by the feature combination used. ANOVAs followed by TukeyHSD tests indicated that the FULL and NoSAL visualizations were each rated as having more relevant annotations than the NoCOS visualizations (μ_{FULL} = 3.6, μ_{NoSAL} = 3.6, μ_{NoCOS} = 2.6, $F(2, 309) = 54$, $p < 0.001$, both $p_{adj} = 0.000$). Similarly, ratings of visual saliency (interestingness) were highest with the FULL and NoCOS visualizations and lowest with the NoSAL visualization (μ_{FULL} = 4.2, μ_{COS} = 4.0, μ_{NoSAL} = 2.3; $F(2, 309) = 149$, $p < 0.001$, both $p_{adj} = 0.000$). Overall ratings were highest for FULL maps (μ_{FULL} = 3.4, μ_{NoSAL} = 2.8, μ_{NoCOS} = 2.6, $F(2, 309) = 11$, $p < 0.001$, both $p_{adj} = 0.001$). We conclude that both features (cosine similarity and Moran's *I*) benefit map usefulness.

The Pearson's correlation between Moran's I and saliency ratings overall was fairly high (0.67, vs. -0.26 and 0.08 with annotation, overall ratings). Our randomized data method may have produced visualizations with lower Moran's I values than might occur in real data with little spatial autocorrelation. Still, the correlation between the measure and map interestingness is encouraging, as we know of no prior evaluative work concerning Moran's I.

As the PMI for labeled variables and the annotation count varied between maps (*min*: 2, *max:* 7), we confirmed that these factors did not impact ratings. We found no evidence of an association between ratings and annotation count (all Pearson correlation values > -0.05 and < 0.12), nor between PMI and ratings (all Pearson correlation values > -0. 51 and < 0.05). We expect the lack of correlation for PMI to have resulted from the small range of high PMI variables that we included to avoid confounds from PMI differences.

DISCUSSION

Limitations and Challenges

The NewsViews' pipeline displays various dependencies that directly impact the quality of the generated visualizations. The variables that have the highest PMI scores with article text (relative to other possible variables) constrain the final visualization that the system generates. Achieving a database that is broad, recent, and accurate in its data is thus critical. These properties can be strengthened in future iterations via the automation of a table crawler (as in [4]) which could simplify updating the database with newly published data from reputable sources. Further pipeline improvements could address challenges associated with generating maps for articles that provide explicit statistics, such as by citing the levels of a variable at different locations. We experimented with simple heuristics for ensuring article-map data alignment such as by looking for exact matches with a variable name with surrounding statistics in the text, and checking that these values aligned with the generated thematic map. There is room for more sophisticated methods for ensuring article-map alignment for these "stats-heavy" articles.

Pipeline Generalization and Future Work

The NewsViews' approach assumes that the content of a news article implies a particular set of *data comparisons*, corresponding to a view on a table comprised of locations and times (rows) and variables (columns). Our high level criteria of variable-to-article similarity (e.g., relevance), database search, and data assignment methods should generalize to other data types. Our demonstration of using PMI to capture variable to article text relevance, and our topic extraction techniques should apply to other visualization types as well. Other high level criteria in the NewsViews pipeline, such as modeling visual interestingness, may require adapting the specific methods used (see, e.g., [12]).

Incorporating additional data extraction algorithms allows the NewsViews pipeline to be extended to visualizations that address different implied comparisons, including temporal, temporal-geographic, and limited group-category comparisons. Figure 4, bottom, displays a line graph generated for an article focused on temporal comparisons, achieved using HeidelTime [30] to produce a "temporal score" capturing the relative degree to which article text focuses on comparisons between an indicator at different times. A "location score" can likewise be achieved using techniques currently implemented in NewsViews. For example, calculating the number of locations mentioned in relation to the anticipated familiarity of those locations to users (inferred from population data, for example). This enables prioritizing map creation where locations are less likely to be known, and where a map's locational affordances are more useful.

CONCLUSION

We present an automated pipeline for generating relevant geovisualizations given context articles. By mining a context article locations and topic, NewsViews can select from a range of data views. We demonstrate that the system is able to generate visualizations that are both "interesting" and relevant. By automating visualization construction, our system can be applied in resource constrained environments that nonetheless have a huge corpus of articles.

ACKNOWLEDGEMENTS

This work is partially supported by NSF Grant SES-1131500.

REFERENCES

1. ArcGIS Resource Center. Map the Data. *ArcGIS Resource Center*, 2013. http://bit.ly/18bKiEw
2. Bouma, Gerlof. Normalized (Pointwise) Mutual Information in Collocation Extraction. *Proc. of the Biennial GSCL Conference*, (2009), 31-40.
3. Brewer, C. Guidelines for Use of the Perceptual Dimensions of Color for Mapping and Visualization, *Proc. SPIE '94* vol. 2171 (1994), 54-63.
4. Cafarella, M.J., Halevy, A., Wang, D.Z., Wu, E., and Zhang, Y. WebTables: Exploring the Power of Tables on the Web. *Proc. VLDB Endow. 1*, 1 (2008), 538–549.
5. Cairo, A. *The Functional Art: An Introduction to Information Graphics and Visualization*. New Riders, 2012.
6. Chang, K. *Introduction to Geographic Information Systems*. McGraw-Hill, New York, NY, 2012.
7. Franklin, B. Key *Concepts in Journalism Studies*. SAGE Publications, London; Thousand Oaks, 2005.
8. Gabrilovich, E., Dumais, S., and Horvitz, E. Newsjunkie: Providing Personalized Newsfeeds via Analysis of Information Novelty. *Proc. of WWW* (2004), 482–490.
9. Griffin, J. and Stevenson, R. The Effectiveness of Locator Maps in Increasing Reader Understanding of the Geography of Foreign News. *Journalism & Mass Commun. Quarterly* 71 (4) (1994), 937-946.
10. Hecht, B., Carton, S.H., Quaderi, M., et al. Explanatory Semantic Relatedness and Explicit Spatialization for Exploratory Search. *Proc. SIGIR '12*, (2012), 415–424.
11. Heer, J., Viegas, F.B., and Wattenberg, M. Voyagers and voyeurs: Supporting Asynchronous Collaborative Visualization. *Commun. ACM 52*, 1 (2009), 87–97.
12. Hullman, J., Diakopoulos, N., and Adar, E. Contextifier: Automatic Generation of Annotated Stock Visualizations. *Proc. CHI '13* (2013), 2707–2716.
13. Hullman, J. and Diakopoulos, N. Visualization Rhetoric: Framing Effects in Narrative Visualization. *IEEE TVCG 17*, 12 (2011), 2231–2240.
14. Jenks, G. and Caspall, F.C. Error on Choroplethic Maps: Definition, Measurement, Reduction. *Annals of the Assoc. of Amer. Geographers* 61, 2 (1971), 217–244
15. Kandogan, E. Just-in-time Annotation of Clusters, Outliers, and Trends in Point-Based Data Visualizations. *IEEE VAST '12*, (2012), 73-82.
16. Kong, N. and Agrawala, M. Perceptual Interpretation of Ink Annotations on Line Charts. *Proc. UIST '09* (2009), 233–236.
17. Kong, N. and Agrawala, M. Graphical Overlays: Using Layered Elements to Aid Chart Reading. *IEEE TVCG 18*, 12 (2012), 2631–2638.
18. Ratinov, L., Roth, D., Downey, D., and Anderson, M. Local and Global Algorithms for Disambiguation to Wikipedia. *ACL* (2011).
19. Leonhardt, D., Bostock, M., Carter, S., et al. In Climbing Income Ladder, Location Matters - NYTimes.com. *The New York Times*, 2013. http://nyti.ms/1dPGWbN.
20. Lin, S., Fortuna, J., Kulkarni, C., Stone, M., and Heer, J. Selecting Semantically-Resonant Colors for Data Visualization.
21. Marcus, A., Wu, E., and Madden, S. Data In Context: Aiding News Consumers while Taming Dataspaces. *DBCrowd 2013*, (2013), 47.
22. Narrative Science. http://narrativescience.com/.
23. OpenCalais, http://www.opencalais.com/
24. Sandhaus, E. The New York Times Annotated Corpus. *LDC*, 2008. http://bit.ly/1a4ObME.
25. Segel, E. and Heer, J. Narrative Visualization: Telling Stories with Data. *IEEE TVCG 16*, 6 (2010), 1139–1148.
26. Shahaf, D., Guestrin, C., and Horvitz, E. Metro maps of information. *SIGWEB Newsl.*, 4, (2013), 1–4.
27. Shahaf, D. and Guestrin, C. Connecting Two (or Less) Dots: Discovering Structure in News Articles. *KDD 5*, 4 (2012), 1–24.
28. Shneiderman, B. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. *IEEE Symp. on Visual Languages '96*. (1996), 336–343.
29. Slocum, T.A. and Slocum, T.A. Thematic Cartography and Geovisualization. Pearson Prentice Hall, Upper Saddle River, NJ, 2009.
30. Strötgen, J. and Gertz, M. HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions. *Proc. IWSE '10* (2010), 321–324.
31. Tenore, M.J. Explainer Maps Locate, Contextualize and Localize News from Libya, Japan. *Poynter*, 2011. http://www.poynter.org/how-tos/newsgathering-storytelling/125206/explainer-maps-locate-contextualize-and-localize-news-from-libya-japan/.
32. U.S. Census Bureau. Statistical Abstract of the United States: 2012 (131st ed.) Washington, DC, 2011; http://www.census.gov/compendia/statab/.