Frankenplace: Interactive Thematic Mapping for Ad Hoc Exploratory Search

Benjamin Adams
Centre for eResearch
Dept. of Computer Science
The University of Auckland
Auckland, New Zealand
b.adams@auckland.ac.nz

Grant McKenzie
Department of Geography
University of California,
Santa Barbara
Santa Barbara, CA, USA
grant.mckenzie@geog.ucsb.edu

Mark Gahegan
Centre for eResearch
Dept. of Computer Science
The University of Auckland
Auckland, New Zealand
m.gahegan@auckland.ac.nz

ABSTRACT

Ad hoc keyword search engines built using modern information retrieval methods do a good job of handling fine-grained queries. However, they perform poorly at facilitating spatial and spatially-embedded thematic exploration of the results, despite the fact that many queries, e.g. *civil war*, refer to different documents and topics in different places. This is not for lack of data: geographic information, such as place names, events, and coordinates are common in unstructured document collections on the web. The associations between geographic and thematic contents in these documents can provide a rich groundwork to organize information for exploratory research. In this paper we describe the architecture of an interactive thematic map search engine, Frankenplace, designed to facilitate document exploration at the intersection of theme and place. The map interface enables a user to zoom the geographic context of their query in and out, and quickly explore through thousands of search results in a meaningful way. And by combining topic models with geographically contextualized search results, users can discover related topics based on geographic context. Frankenplace utilizes a novel indexing method called geoboost for boosting terms associated with cells on a discrete global grid. The resulting index factors in the geographic scale of the place or feature mentioned in related text, the relative textual scope of the place reference, and the overall importance of the containing document in the document network. The system is currently indexed with over 5 million documents from the web, including the English Wikipedia and online travel blog entries. We demonstrate that Frankenplace can support four distinct types of exploratory search tasks while being adaptive to scale and location of interest.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: General; H.2.8 [**Database Management**]: Database Applications—*Spatial Databases and GIS*

WWW 2015, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3469-3/15/05.
http://dx.doi.org/10.1145/2736277.2741137.

General Terms

Measurement, Theory

Keywords

Geographic search; interactive search; information retrieval; information visualization; visual analytics; exploratory search

1. INTRODUCTION

Geographic location is playing an increasingly important role in information retrieval: to find places of interest that match a user's query and for personalizing search based on user location [44, 10]. Most of this work focuses on how to improve localized search results for a standard presentation like the top-N search page [16, 8, 4]. In this paper we propose that geographic references in text also provide a powerful way to organize and explore ad hoc information via an interactive thematic map interface. Ad hoc information retrieval is the task of finding relevant documents from a corpus given an unconstrained keyword query. Ad hoc retrieval implies that users will perform many different kinds of searches and for many different purposes, not only to find a specific document but for more exploratory reasons as well. Faceted browsing for ad hoc web search faces certain challenges not least due to the difficulty in coming up with a categorization scheme that is broadly useful [7]. This in part is why the non-faceted keyword search model exemplified by Google and other modern search engines supplanted the hierarchical category-based faceted browsing that companies like Yahoo! developed in the late 90's. We argue that there is a case to make again for spatial and temporal faceting for ad hoc search, because 1) geographic space is a common ordering principle in human discourse that is broadly applicable to a number of different kinds of searches, and 2) a massive number of online documents have spatial and temporal semantics that allow us to interpret their content along these dimensions and the technical capability to model these semantics from unstructured content has by now been developed.

Cartographers have known for centuries the power of organizing thematic information by geographic context and visualizing it on a map. This allows one to understand the spatial dynamics of a topic within a structure that has strong reference points in human experience (i.e., the places represented on the map). At the same time it tells us important information about the places, which can be used for their comparison (similarity, groupings, etc.). Recently, we have begun to see examples of thematic mapping interfaces for posts from social media applications like Twitter.[1] However, in these cases the georeferencing is based on simple point data provided by the service and there is very little work on using these interfaces for anything beyond simple visualization. Partly, this is because of the limited value in the retrieval of individual tweets that match the query versus the information gained from viewing the aggregate visualization from millions of tweets. The visualization alone is a useful product for understanding the query as it relates to place, but when coupled with ad hoc retrieval it becomes a way to geographically contextualize the documents for the user. For many corpora, such as web documents, there is real value in using the geographic context as a means to discover, organize, and interactively visualize the documents related to a search query. One obvious benefit of such a method is that instead of merely returning a list of the top-N matching documents, a thematic map enables geographic exploration of thousands of results.

In this paper we discuss the implementation of a system for geographic exploratory search of Wikipedia articles and online travel blog entries, called Frankenplace[2] (see Figure 1). In the design of such a system we are presented with the question of how to rank the relative importance of documents in a way that is sensitive to the geographic scale the user is interested in. That is, given that we know a place is associated with a document, how should the scale of that place reference inform how the document content is reflected in a thematic map at a given zoom level? For example, if a corpus contains articles about California (but without references to more localized places), but the user is zoomed into the local San Francisco area, should those California-related articles still be rendered in some form on the resulting map because they spatially intersect with the map window, or should only locally-referenced content be displayed? Utilizing a hierarchical discrete global grid, we introduce a text indexing method that adjusts the importance of documents based on the geographies of referenced places, so that search results are attuned to zoom level. In addition, apart from administrative regions, most popular map-based search tools (e.g., Google Maps and Bing Maps) focus on retrieving point-based geographic features (i.e., points of interest), even though visually representing *regions* or *areas of interest* is of significant value to the spatial search community. Our global grid approach combined with kernel density estimation provides a mechanism to visualize search results in a variety of ways, including as regions and surfaces.

It is helpful to consider a use case at this point. An economic historian is interested in exploring the evolution and history of ghost towns in the United States as they relate to economic changes, geographic environment, and cultural events such as the gold rush. At a preliminary stage of research this historian will first want to gather information about ghost towns in geographic context. A simple approach is to start using a search engine, such as Google, to explore for information by searching for *ghost town*. However, this search, as shown in Figure 1, returns 14.9 million results; it is unclear how to explore the results geographically. The historian can refine the search by adding geographic terms but this requires knowing in advance which locations to search for. In addition, the Google map interface is designed to emphasize commercial points of interest, and thus the results are not particularly useful for this task. In contrast, by organizing document information spatially and enabling interaction via a map interface, several more useful documents can be uncovered. Furthermore, geographically related topics such as *gold rush* become clear as the user navigates through the system. The quality of the results will, of course, depend on the underlying corpora used to build the index. Continued development to include additional sources, such as primary historical documents, will only improve the ability to perform these kinds of searches.

In this paper we detail the creation of a thematic map exploratory search engine designed to enable this kind of research. Our main novel contributions are as follows:

1. The use of a hierarchical discrete global grid to organize the geographic indexing of text to enable easy integration from multiple sources at multiple zoom levels.

2. A method for boosting terms associated with cells on the discrete global grid that factors in the geographic scale of the place or feature mentioned in related text, the relative textual scope of the place reference, and the overall importance of the term's document in the document network.

3. An interactive thematic map search interface that enables users to quickly browse to refine or expand the scope of their query geographically in an integrated system.

4. A technique to suggest new searches based on the current search results, filtered by a selected grid cell. Topic modeling is used to calculate a topic vector associated with each document when the index is created. Topic vectors are aggregated on the fly at query time providing suggestions that are query and geography specific.

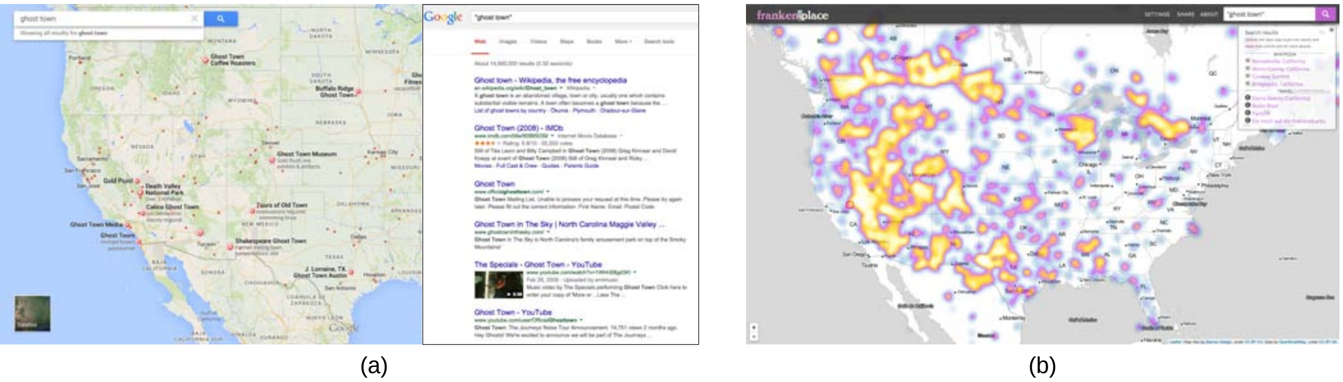
The remainder of the paper is organized as follows. After a brief overview of the system architecture, Section 3 describes our methodology for geographic indexing of documents. Section 4 details the exploratory search interface design and implementation. In Section 5 we discuss some of the novel search tasks that Frankenplace enables with evaluation. Related work and finally our conclusion and directions of future work close out the paper.

2. SYSTEM OVERVIEW

The Frankenplace system architecture is shown in Figure 2. In the backend of the system, geographic data from multiple sources is combined to create a gazetteer that is used to match place names in documents to grid cells on a discrete global grid. Several sources of information are organized into two types of indices that are built to support geo-thematic search (details in Section 3). The first type is an inverted index of terms associated with grid cells, and the second is an index of individual documents and related topics. An exploratory search web client interactively renders search results from these indices to map out prominent places and highlight matching documents based on user selected locations (details in Section 4).

[1] See e.g., http://mapd.csail.mit.edu/tweetmap-desktop

[2] http://frankenplace.com. Note, a very early version of Frankenplace with substantially different architecture and functionality was previously described in [2].



(a) (b)

Figure 1: Comparison of search results for the query *"ghost town"* from (a) a popular search engine, (b) Frankenplace system.

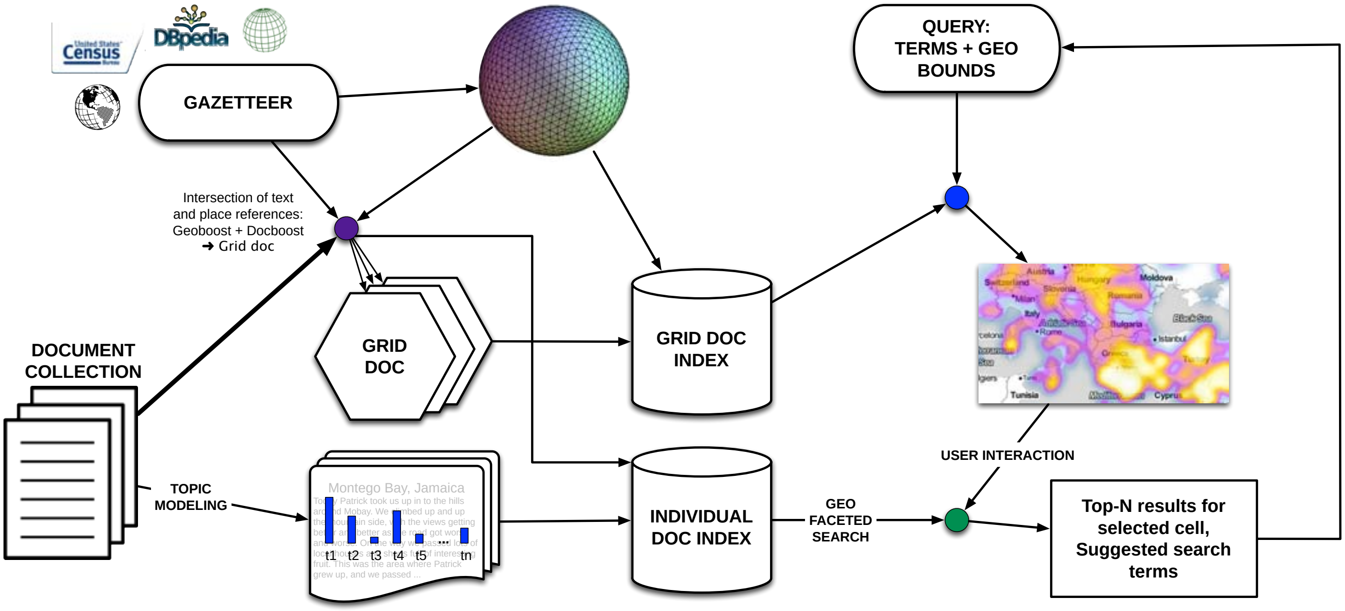


Figure 2: Overview of the Frankenplace system architecture from indexing to interactive search.

3. INDEXING

For a thematic keyword search that identifies relevant places based on the content of documents, one must create an inverted index from terms to places and regions on the Earth. In this section we describe the design of such an index (and related ranking algorithms) when the input data consists of millions of documents that intersect with these places in differing capacities. In a hyperlinked database, statistical network measures such as HITS are very effective at uncovering "important" documents [23]. But here we supplement this with an importance score for document sections based on their relationships to places. These scores are derived from both *overall* and *partial* references to places within a document. An overall reference means the entire document is about a place, and thus all the text contained within is relevant to index that place. An example of this would be a travel blog entry about the city of Tokyo. A partial reference, in contrast, is located within a text but only some subset of the content is related.

3.1 Mapping place names to global grid

Some work has been done on matching geographic information in text to geodesic grids, but without much care over geographic projection and consequent areal distortions [43, 1]. In contrast, in this work we create an information retrieval index that flexibly customizes results for multiple geographic scales (map zoom levels). Thus, it is important to design a system that organizes the information such that it facilitates scale dependency. Digital Earth systems that are designed to operate on multi-thematic, global geographic data sets (such as climate data) use discrete global grids (DGGs) to organize their information in this way [12]. The use of a DGG has the added advantage that search results can potentially be integrated with geographic information from a number of other geophysical and social science data sets, including land use, climate, and demographic data.

A DGG is a partitioning of the entire Earth's surface into equal area (or approximately equal area) grid cells at multiple, interrelated levels of resolution that can be indexed and searched quickly by geographic coordinates [32]. There are several approaches to building DGGs, but one common approach is to start with a polyhedron simplification of the spherical globe (usually an icosahedron) and then hierarchically partition the faces of the polyhedron into triangles, diamonds, or hexagon shaped cells. Depending on the orientation of the initial polyhedron, the transformation function between the polyhedral surfaces and the spherical globe, and the spatial partitioning technique used, several types of DGGs can be generated [36, 17]. In addition, alternative methods exist that do not use a polyhedral simplification, such as HEALPix, which was developed by the astronomical community for mapping star densities [13]. For web mapping (our purposes), a hierarchical triangular discrete global grid has the advantage that at every successive level the number of cells increases by a factor of four–the same as web tiles. As a result, the size of the grid cells can remain constant relative to the map features at all zoom levels.

Several online gazetteers exist that map place names to locations on the Earth, including geonames.org, NGA GEOnet Names Server, Getty Thesaurus of Geographic Names, and quattroshapes, as well as other sources containing this information, such as DBpedia structured data sourced from Wikipedia. In addition, coming out of the digital humanities community there is an increasing number of specialized gazetteers that focus on historical place names [37]. Despite movement toward the use of semantic technologies in gazetteers [22], these gazetteers vary widely in terms of their quality of representation, especially with respect to interoperability, classification, toponyms in multiple languages, and the geometric representation of spatial footprints. Most entries in available gazetteers use point geometry or occasionally simple bounding box representations for the spatial footprint of a place. As a result, in our work we found it was best to align identifiers across gazetteers and then improve the geometry when possible by linking in polygon data from national and regional authorities. For that task we included country, first order administrative region, and natural park polygon data from the North American Cartographic Information Society,[3] and TIGER data from the United States Census Bureau.[4] This gazetteer enrichment remains an ongoing process. The next step is to map each of these gazetteer entries to sets of cells in the discrete global grid. This is done by calculating the spatial intersection between the polygon or point representation of the place and grid cells at each level. The result of this process is a set of mappings from places in the gazetteer to sets of grid cells at multiple levels of the DGG.

3.2 Assigning weighted words to grid cells

We start with a collection of documents and place name entries in our gazetteer and from that we need to assign weight values of words for cells in the discrete global grids. We define two types of scopes for place references in a document: 1) *overall* place references, which are relationships between a place and the entire document; and 2) *partial* place references, which are places that are only related to a local scope of text, such as a sentence, paragraph, or section.

Identifying places in Wikipedia. Wikipedia has a great deal of structured geographic information associated with articles, often in the form of geographic coordinates, which have been extracted from Wikipedia into the DBpedia knowledge graph [25]. Since we want better-than-point representations when possible, it is not sufficient to simply match overall place references to latitude-longitude references on Wikipedia pages, however. Instead multiple sources of evidence are brought together to match articles to the identifiers in our integrated gazetteer. This gives us an excellent start to match places within the text as well. Links within a page that are instantiated in DBpedia as the `dbpedia:place` type are identified as partial place references. In addition, links to `dbpedia:event` pages that also have a georeference are matched. Since it is customary for Wikipedia pages only to link to the first occurrence of a related page, it is necessary to match for additional occurrences and assign links there as well, otherwise we will miss many paragraphs with place references.

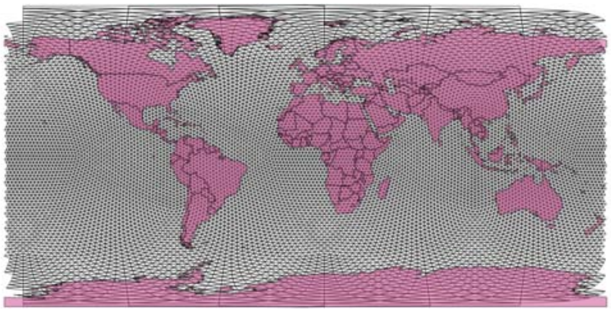


Figure 3: Triangular discrete global grid with approximately equal area cells.

Identifying places in travel blog entries. Travel blog entries come in a variety of formats on the web, so matching overall and partial place references in travel blog entries is more difficult. For this task we leveraged both the CLAVIN geoparser[5] and geonames.org web services to match place names to geonames ids that we could then match to points or areas in our integrated gazetteer. Improving geoparsing methods was not the main subject of this research; existing solutions work fairly well, though there is much room for improvement since false positives and missed place name matches occur often.

Computational sensemaking of how places relate to the discursive and narrative structure of a document remains at the frontier of natural language processing research and is beyond the scope of this research. However, assuming a linear narrative structure in most documents, we make the simplifying assumption that the scope of a partial place reference can be set to a fixed syntactic structure that the place reference falls within. In some sources, such as Wikipedia, the syntactic structure of paragraphs are clearly delineated and work as a useful proxy for text related to a place [20]. Crowdsourced information that has less quality control (such as travel blog entries) will need a more flexible scoping such as a sliding window of words.

3.2.1 Geoboost

Once we have relationships between words (in sections) and place names, it is possible to calculate a *geoboost* value in the range (0,...,1] associated with each word for each grid cell. In this way the same word from a given document can

[3] `http://www.naturalearthdata.com/`

[4] `https://www.census.gov/geo/maps-data/data/tiger.html`

[5] `https://github.com/Berico-Technologies/CLAVIN`

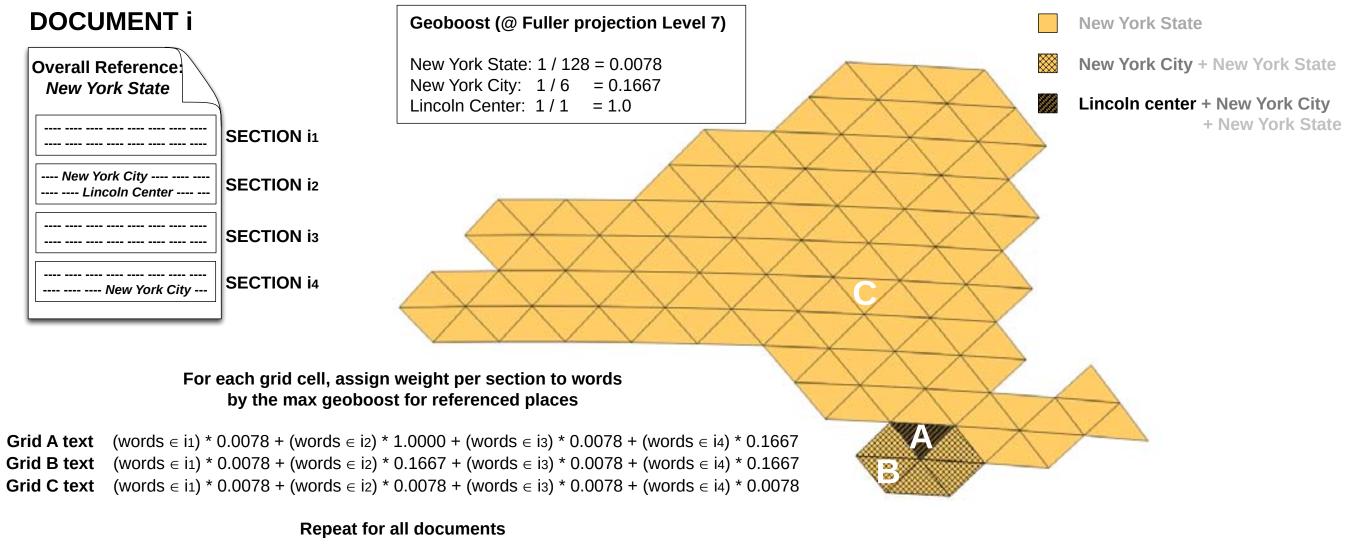

DOCUMENT i
Overall Reference: New York State
---- New York City ---- ---- ---- ---- Lincoln Center ---- ---
---- ---- ---- New York City ---
SECTION i1
SECTION i2
SECTION i3
SECTION i4
Geoboost (@ Fuller projection Level 7)
New York State: 1 / 128 = 0.0078
New York City: 1 / 6 = 0.1667
Lincoln Center: 1 / 1 = 1.0
New York State
New York City + New York State
Lincoln center + New York City + New York State
A
B
C
For each grid cell, assign weight per section to words by the max geoboost for referenced places
Grid A text (words ∈ i1) * 0.0078 + (words ∈ i2) * 1.0000 + (words ∈ i3) * 0.0078 + (words ∈ i4) * 0.1667
Grid B text (words ∈ i1) * 0.0078 + (words ∈ i2) * 0.1667 + (words ∈ i3) * 0.0078 + (words ∈ i4) * 0.1667
Grid C text (words ∈ i1) * 0.0078 + (words ∈ i2) * 0.0078 + (words ∈ i3) * 0.0078 + (words ∈ i4) * 0.0078
Repeat for all documents

Figure 4: Example of how words are assigned with geoboosts from a sample document containing references to New York State, New York City, and Lincoln Center.

contribute with different weights to different grid cells. The *geoboost* (Equation 1) is a function of the number of grid cells covered by the place references associated with the word at a given resolution. Let $w_{i,j,k}$ be defined as word k from section j of document i. Let $\sigma(w_{i,j,k})$ be the section of word $w_{i,j,k}$, $pp(\sigma(w_{i,j,k}))$ be equal to the set of partial place references in the section, and $c(l, pp(\sigma(w_{i,j,k}))$ be the number of grid cells covered by the partial place at grid level l.

$$geoboost(g_{l/id}, w_{i,j,k}) = \frac{1}{min(c(pp(\sigma(w_{i,j,k}))))} \qquad (1)$$

The rationale for geoboost is that more fine-grained place references should override coarse-grained place references when it comes to assigning importance of text to a grid cell. If a place reference only falls within one grid cell at a given scale, then the surrounding text should be weighted highly for that grid cell and less so for other grid cells that are covered by other larger area references in the same section. Figure 4 illustrates how the assignment of geoboosts is made with a simple example using three place references: New York State, New York City, and Lincoln Center. By tying the geoboost value to the number of grid cells matching the area of the place at a specific level, we automatically take into account the geographic scale at which the user is conducting the information retrieval task. For example, when using a coarse-grained grid (i.e., searching a zoomed out map) then text which references administrative regions, cities, and other areal spatial units will be as important as references to individual points of interest. When zoomed in and using a finer-grained grid, text associated with point of interest features will contribute more to a grid cell.

As is common for information retrieval in networked document collections, a relative importance weight for the entire document can also be calculated using HITS, PageRank, or similar algorithms [23, 30, 26]. In order to differentiate from the geoboost, we refer to the document weight in the network as the *docboost*. Combining these two values gives us an *composite boost value* for an individual word that has been mapped to a grid cell, equal to $docboost^m * geoboost^n$ where m and n are exponential scaling factors used to increase or decrease the relative weight of docboost versus geoboost. Finally, we can create a *grid document* that aggregates all the words with their boost values in all the sections that have place references intersecting with the cell. The *grid document* is a summary of all the content written about that grid area in such a way that text is considered to matter more or less depending on the scale of place references and importance of the document. The advantage of this approach is that the resulting *grid document* has the same form as a single document, and thus can serve as input into a number of compatible information retrieval indexing algorithms [26].

3.2.2 Probabilistic index

We can anticipate spatial heterogeneity in terms of the distribution of source material. That is, certain places (such as London, UK in the English Wikipedia) will have orders of magnitude more words indexed against them than other places that are represented in the source material. In our experiments, the distribution of word counts for grid documents tends to follow a log normal distribution (see Figure 5). For ranking this creates a problem similar to the case of matching a query against a corpus with documents of widely differing lengths, so it is necessary that document length normalization take this into account.

Although the grid document can serve as input into any indexing algorithm, because the grid documents are themselves aggregate documents, we need a model of information retrieval that captures the informative content of terms with respect to the structure of the documents and overall corpus. This is better handled by probabilistic models of information retrieval than by standard TF/IDF related measures [31, 3, 26]. Divergence-from-randomness (DFR) is one such model incorporated in the Terrier IR platform that uses the Bose-Einstein distribution as the model to test the divergence of word mentions from that expected from a random document [3]. In Frankenplace we use the slightly simpler information model derivative of DFR as described in

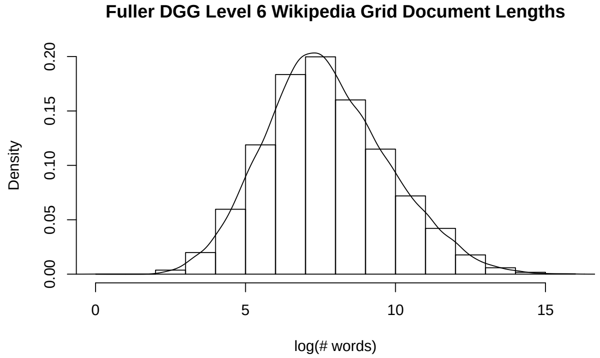


Figure 5: Log plot of word counts for the English Wikipedia grid documents for the Fuller level 6 discrete global grid.

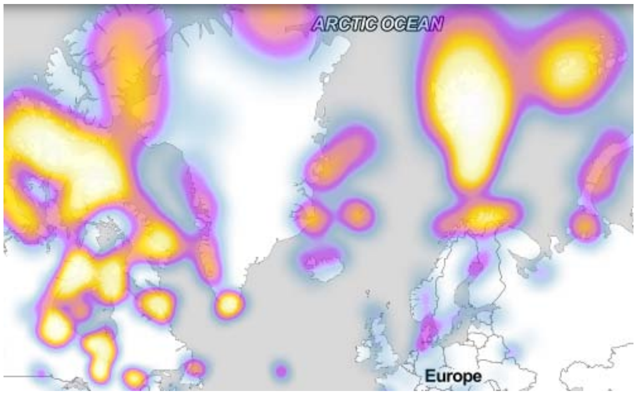


Figure 6: Heatmap that demonstrates the increasing bandwidth of the kernel density approaching latitudes farther north. (*polar bear* search)

[9], with word counts adjusted in incorporate the *composite boost values*.

3.3 Indexing individual documents

We are interested not only in searching for grid cells that match a query but also want to find the top documents that both match a grid cell and the query terms. This is necessary in order to make a fully ad hoc document search engine, otherwise we are left with matching the query only at the aggregate level. The solution is to build a parallel index for all the documents, faceted by grid cell, so that queries can be filtered to match a single or set of grid cells. As for the grid cell index, we boost the terms in the individual document by geoboost and docboost. However, we also want an additional damping effect to take into account cases where not all words in the document are associated with the grid cell (as will be the case in most partial place references). This damping factor, d, is the ratio shown in Equation 2.

$$d = \frac{\#\ words\ in\ the\ matching\ sections}{total\ number\ of\ words\ in\ the\ document} \qquad (2)$$

4. BUILDING A THEMATIC MAP EXPLORATORY SEARCH ENGINE

In this section we describe the design of an interactive map search interface once indices have been created for grid cell documents (at multiple zoom levels) as well as for individual documents. A thematic map can be viewed as a specific type of information visualization, and from a user interface design and engineering perspective, building a thematic map keyword-based web search is more challenging than a traditional top-N search result page. The design choices are complex, because the manner in which ranking scores are visually rendered on the interactive map will have a bearing on user interpretation.

When a user submits a query, there are two related types of search results that need to be rendered. The first is a global view of the results that shows what grid cells on the Earth best match the query. This global view is a map of the search results over geographic space. The second result provides a localized view of the results by showing the top documents for the query faceted by the grid cell that is currently selected by the user. These two related results should be consistent with one another, but they have different purposes. The first search (the global map) is a ranking of *places* that match the query, and the second search is a ranking of *documents*.

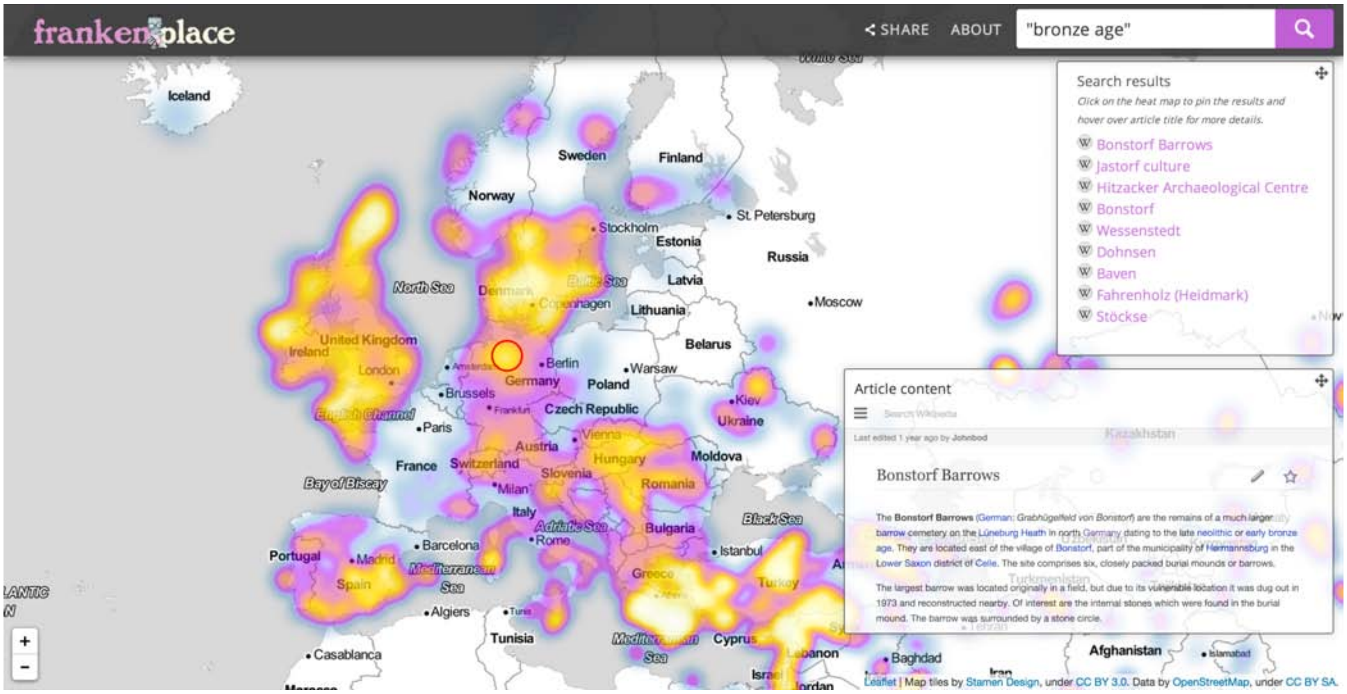
A highly ranked grid cell should correspond in some way to a good match for the query based on an aggregation of documents for that cell, but it is not necessary that a highly scored grid cell (i.e., place) means there should necessarily be many document results. For example, a single document with high *geoboost* and *docboost* that ranks very highly for a given query (e.g., many instances of the query term in the document) should be reflected in a higher ranking score for the grid cell that the document falls within, whereas a grid cell that has many documents that match the query but which are all very low on *geoboost* and *docboost* might correspond to a grid cell with very low score. The overall boosts that have been attached to terms in the index are designed to handle these cases.

4.1 Map visualization

One option is to render the search scores for the grid cells using a choropleth map[6], so that scores are mapped to an ordinal range of colors. Care must be taken to use a color scheme that accurately renders the relative scores of cells in a manner that is cognitively valid [6]. One solution is to use relatively small number of color classes (7 or fewer) that cluster the scores via a method such as Jenks natural breaks [21]. However, one drawback of the choropleth map is that the grid cells are arbitrarily discrete and convey a degree of certainty in the result that may be unmerited, and changing the grid can create very different results (because of the modifiable areal unit problem) [11].

With this in mind, we instead use a two-dimensional kernel density function with a bandwidth based on zoom level to generate a smooth surface of the search results over the map (i.e., a heat map) [38]. The input to the kernel density function is a set of points corresponding to the centroids of the grid cells, with count values set to the respective search score. In order to calculate the kernel density efficiently on the client-side of the application, a simple radial-symmetric triangular kernel is used [34]. An additional challenge is

[6] A map with regions that are colored or shaded to represent an associated attribute value.



Search results

Click on the heat map to pin the results and hover over article title for more details.

- Bonstorf Barrows
- Jastorf culture
- Hitzacker Archaeological Centre
- Bonstorf
- Wessenstedt
- Dohnsen
- Baven
- Fahrenholz (Heidmark)
- Stöckse

Article content

Search Wikipedia

Last edited 1 year ago by Johnbod

Bonstorf Barrows

The **Bonstorf Barrows** (German: *Grabhügelfeld von Bonstorf*) are the remains of a much larger barrow cemetery on the Lüneburg Heath in north Germany dating to the late neolithic or early bronze age. They are located east of the village of Bonstorf, part of the municipality of Hermannsburg in the Lower Saxon district of Celle. The site comprises six, closely packed burial mounds or barrows.

The largest barrow was located originally in a field, but due to its vulnerable location it was dug out in 1973 and reconstructed nearby. Of interest are the internal stones which were found in the burial mound. The barrow was surrounded by a stone circle.

Figure 7: Search result for *bronze age* using Frankenplace. This screenshot shows the global result as a heat map over Europe, and the local result pinned to an area of northern Germany (red circle). The top documents for the local result are shown in a window on the right side of the screen, and a preview of one selected document is shown at the bottom.

presented when generating a heat map in the standard web Mercator projection from data that is based on an equal area grid, because the centroids of the grid cells are farther apart in the projected space as one gets closer to the poles.

Our solution is to vary the bandwidth, h, in the kernel density function based on a bandwidth scale factor equal to the secant of the latitude (Equation 3).

$$h' = h \cdot \sec(\phi) \tag{3}$$

This creates an *approximation* of a heat map generated in an equal area space and then projected to web mercator. Figure 6 shows an example of a heat map that significantly increases in bandwidth as it approaches the north polar region.

The visualization and user interface shown in Figure 7 is built on open source mapping software (`http://leafletjs.com/`) and uses HTML5 SVG. The heat map creates a visual overlay on the map but the interaction events are handled by creating a transparent SVG div of the original grid cell polygons overlaying the map. This way, click and mouseover events can be quickly mapped to grid cell identifiers that are used to facet search on the individual document index.

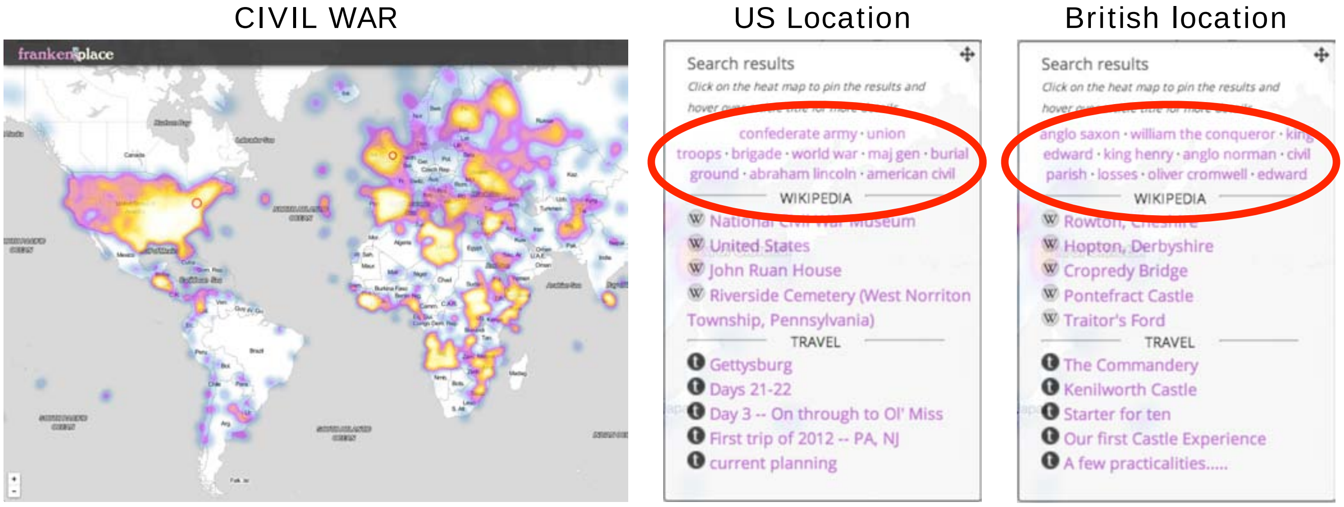
4.2 Exploring by geography and theme

We have established a means to geographically explore the results of a query, but to better enable exploratory search we also need a means to move through the thematic space as well. Probabilistic topic modeling provides a data-driven way to discover the topics in a large corpus of documents, where topics are multinomial distributions over words in the corpus. Latent Dirichlet allocation (LDA) is the simplest and most commonly used generative Bayesian probabilistic topic model [5]. As a pre-processing step during the index building, we run a variant of LDA with β hyper parameter estimation on the Wikipedia corpus to get a fixed number of topics [41]. In addition, this gives us a distribution over topics (i.e., a topic vector) for each document in the corpus. This topic vector is stored in the index as an array associated with each document identifier, so that it can be retrieved at query time.

When a search is made to match documents for a grid cell, the topic vectors are returned along with the ranking score. An aggregate topic vector is calculated at query time equal to the weighted sum of all the vectors from the top n matching documents, where the weight is set to the score from the search algorithm. Let $\mathbf{t}_i$ be the topic vector for document i and s_i be the search score. The aggregate topic vector $\mathbf{a}$ is calculated in Equation 4.

$$\mathbf{a} = \sum_{i}^{n} s_i \mathbf{t}_i \tag{4}$$

By selecting the topics with highest probabilities in $\mathbf{a}$, we can identify important related topics as defined by the original search query and the geography of the grid cell. The most common words from these contextually-related topics are then used as suggestions for new searches in the application. Figure 8 demonstrates this with two different results for the query *civil war* for locations in the United States and England, with very different suggested searches. The grid cell in the United States has related searches *confederate army* and *union troops*, whereas the grid cell in England has *anglo saxon* and *oliver cromwell.* Because data-driven topic models are used, rather than a hierarchical category structure that is pre-built from a global perspective, the se-



Search results

Click on the heat map to pin the results and hover over the title for more details

confederate army · union troops · brigade · world war · maj gen · burial ground · abraham lincoln · american civil

WIKIPEDIA

- National Civil War Museum
- United States
- John Ruan House
- Riverside Cemetery (West Norriton Township, Pennsylvania)

TRAVEL

- Gettysburg
- Days 21-22
- Day 3 -- On through to Ol' Miss
- First trip of 2012 -- PA, NJ
- current planning

Search results

Click on the heat map to pin the results and hover over the title for more details

anglo saxon · william the conqueror · king edward · king henry · anglo norman · civil parish · losses · oliver cromwell · edward

WIKIPEDIA

- Rowton, Cheshire
- Hopton, Derbyshire
- Cropredy Bridge
- Pontefract Castle
- Traitor's Ford

TRAVEL

- The Commandery
- Kenilworth Castle
- Starter for ten
- Our first Castle Experience
- A few practicalities.....

Figure 8: This figure shows how related documents and suggested searches reflect the intersection of geography and the user provided query. The search results, including suggestions for new searches, for the query *"civil war"* are shown for two locations in Pennsylvania, USA and England.

mantic heterogeneity of search terms based on geography is better captured.

With the addition of the related topics into the index the user now has the ability to do exploratory search, zooming in and out in scope both geographically and thematically. This completes the loop in Figure 2, enabling the system to not only work for user-prompted information retrieval but also providing a recommender system to further explore the corpus through the orthogonal lenses of geography and thematic topics.

5. EVALUATION

A live version of the Frankenplace system is available online, which has allowed us to do preliminary observation of exploratory search behavior with the system in the wild. Server logs from 1,697 users give us insight into the kinds of queries that users are performing and how they interact with the system. On average individual users made requests for multiple query plus geographic bounding box combinations per session (mean: 11.4, median: 7). These users searched 4,982 unique terms covering a broad range of topics. Figure 9 shows a sample of the range of query topics, including people, geographic feature types, historical eras and trends. A non-exhaustive taxonomy of search tasks that users perform with the system are listed below.

1. Geographic search of a topic. The interactive thematic map can be used to explore search results via hierarchical geographic facets: by interpreting the overall heat map, zooming in and out, panning from one location to another, and moving the mouse over grid cells to examine the search results at different locations. The search logs indicate that users are performing all of these actions with the system. The number of grid cells the user examines to view search results is mean 26.7, median 13, which indicates that on average users are examining search results over a range of geographic areas. Figure 10 shows the trajectory of mouse movements made by a sample user who is geographically-refining a query for *ski.* Examining users' geographic foci of attention for different queries is potentially a rich source of data for user modeling and predictive analytics.

2. Thematic zooming and panning based on geography. In this case a user begins with a broad topic such as *history* or *civil war* and uses the interactive map to suggest related searches that are geographically contextualized. This allows the user to zoom in the query thematically. In addition, associations based on the intersection of a narrow topic and geography enables thematic zooming out (or horizontal thematic stepping) as well.

3. Comparing corpora. By comparing the maps from different corpora for the same search, it is possible to uncover interesting differences in how a topic is covered by a community. Frankenplace enables this in the current version with a slider that adjusts the weighting of the 2 corpora in the heat map visualization. For example, in Figure 11 it shows how *archaeology* in the context of tourism is most dominant in travel blog entries about Greece, but in Wikipedia there are entries all over the world, including many places where people do not travel and visit archaeological sites.

4. Searching for related places When a user does a search for a place name, as expected from the first law of geography, nearby places are also highlighted on the map [40]. However, connections to other places far away in physical space can also be discovered, reflecting the important relational and networked aspects of places [15]. As an example, Figure 12 demonstrates references to London in articles georeferenced in Italy.

6. RELATED WORK

Although the top-N search result page is the most prevalent way of presenting document-based information retrieval results, there have been a number of research studies looking at alternative approaches to visualizing search results. An early system called Envision visualized various characteristics of the results, including the categories that the document falls within [29]. Schneiderman et al. [33] highlighted the limitations of presenting search results in a top-N list and proposed a system of hierarchical axes, called hieraxes,

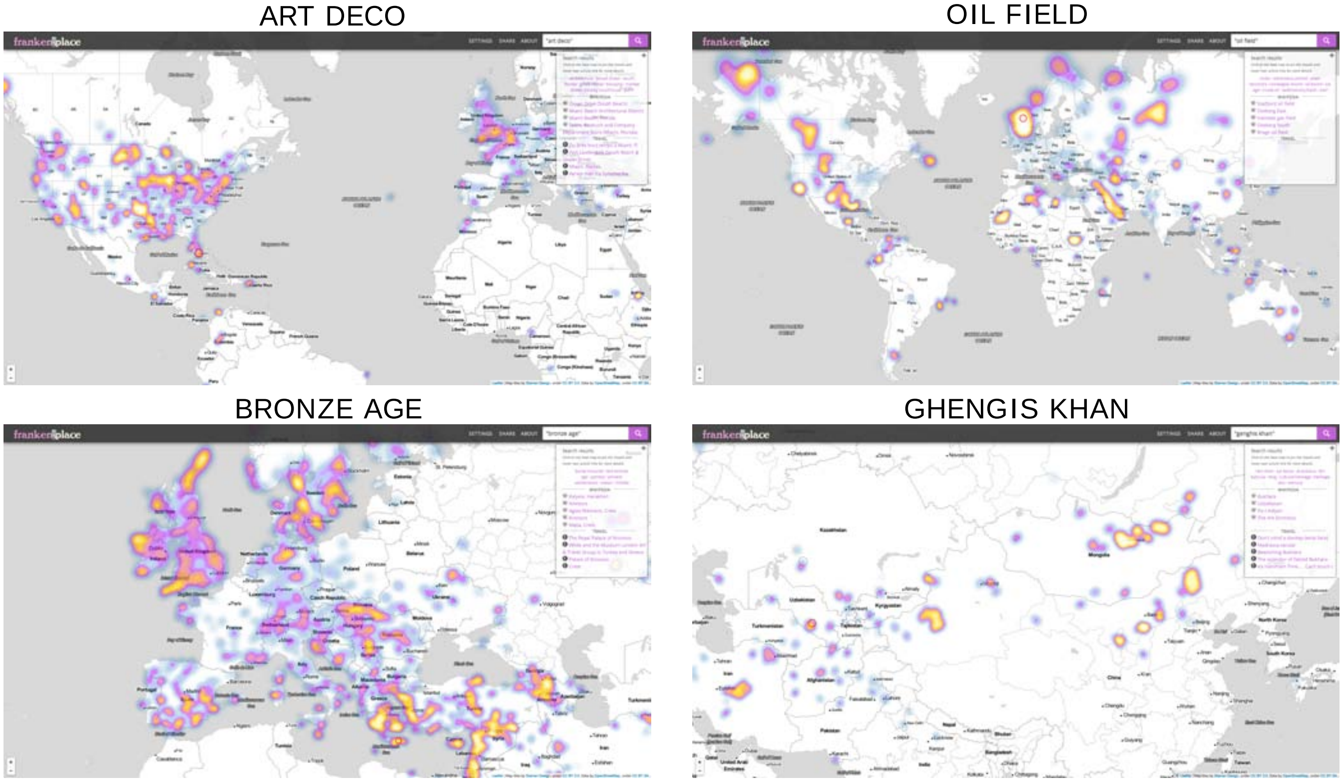


Figure 9: Example searches that users have made.

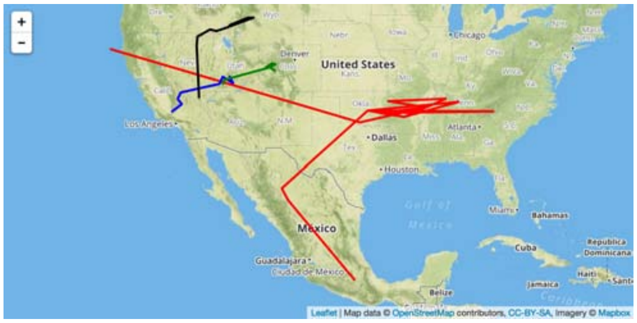


Figure 10: Trajectory of grid cells that the user's mouse passed over while searching for the query *ski*. The red line is map zoom level 3 and shows the user exploring from Mexico with detailed interest in the Ozarks in the southern US (an area not generally well known for skiing) and then the west. The blue, green, and black lines show successive zooming to levels 4, 5, and 6 respectively, as the user focuses the search on the Rocky Mountains.

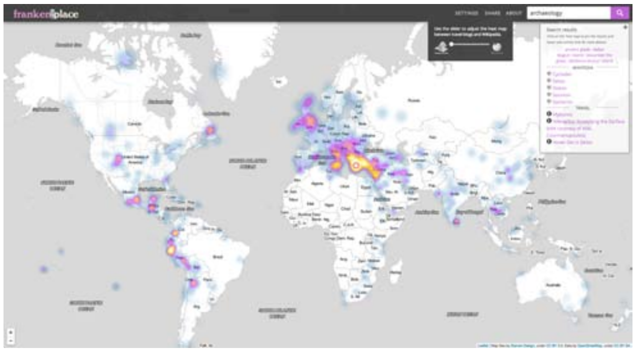
to organize information along two-dimensional axes based on user-defined variables while also letting users zoom in and out based on hierarchical categories. Chen and Dumais [7] suggested the use of support vector machines to automatically hierarchically categorize search results for better presentation. More recently, data driven visualizations of search results based on topic modeling have become popular [28, 18].

Exploratory search differs from the standard lookup model in information retrieval [42]. Whereas the lookup task is concerned with one-time retrieval of facts or information objects such as web pages, the goal of exploratory search is to learn and investigate using systems that support the tasks of knowledge acquisition, synthesis, and discovery [27]. In other words, the goal of exploratory search is to assist human learning, and search tasks are more integrated with browsing activities in an iterative manner, because the results of the information retrieval task lead to new queries [42]. An empirical user study of people doing exploratory search tasks using faceted search interfaces showed that the facets are important informational cues to users [24].

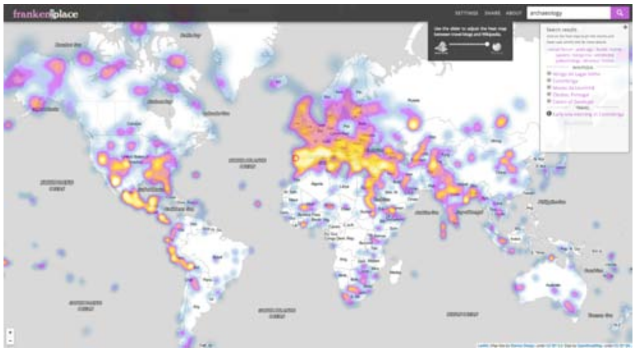
The notion of using space and time dimensions to organize document collections has been advanced in previous work. An early map-based interface for exploring document collections based on historical events was proposed as part of the Perseus Digital Library project [35]. Tezuka, et al. [39] argued that better integration of web search systems with the data models and analytical methods from geographic information systems will enable new search result presentation methods and facilitate knowledge discovery through geographic aggregation. The use of spatialization to enhance exploratory search is the basis for the Atlasify system, which uses a number of spatial layouts (including periodic tables and congressional seating charts) to contextualize search results [19].

7. CONCLUSIONS AND FUTURE WORK

In this paper we presented an innovative system for exploratory search at the intersection geography and theme.



(a)

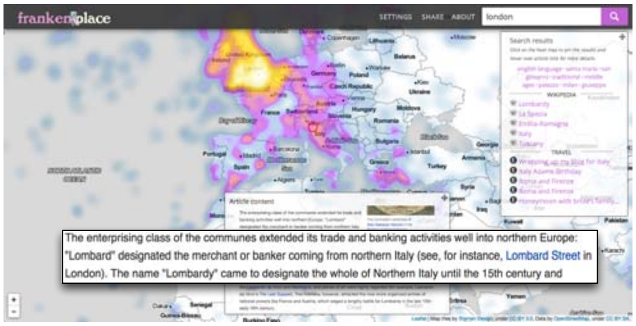


(b)

Figure 11: Maps for *archaeology* for (a) travel blog entries and (b) Wikipedia.

We described a novel approach to indexing using a discrete global grid and term boosting that is geographic scale-dependent. We introduced Frankenplace, a prototype thematic map ad hoc search engine for exploratory search, that utilizes this indexing method on two corpora, and demonstrated how, using topic modeling, it can be used to explore information by zooming in and out both geographically and thematically. As the system continues to be developed, adding new content and features (such as temporal faceting) while adding new users, we anticipate being able to investigate a host of research related to exploratory search with thematic maps. With Frankenplace, we demonstrated we can support four distinct types of exploratory search tasks while being adaptive to both scale and location of interest, and since the index is emergent rather than using formally defined semantics it can easily be extended to new corpora. Our next steps in system development will be to add temporal faceting and an API that will enable users to index and search their own data sets.

Moving forward, combining map-based visualization with ad hoc information retrieval creates many interesting research questions at the intersection of spatial cognition and information seeking behavior. Can we develop appropriate cognitive models of information retrieval tasks given a thematic map interface? What roles do background geographic knowledge and visual interpretation of search results play in how people interpret search results and interact with the system for exploratory search? Extensive research on the cognitive modeling of map understanding can inform the visualizations and interaction paradigms of future iterations. Any map-based visualization of search results will undoubtedly lead to interpretations about the places themselves, but as has been well established for decades, comparisons of places based on aggregate measures are prone to problems such as the ecological fallacy [14]. Thus, there is the question of how to reconcile the opaqueness (to the user) of the indexing process in a modern information retrieval system–including the design choices for spatial and temporal organization–with the common tendency to view thematic maps as some kind of "truth" about the places represented.



The enterprising class of the communes extended its trade and banking activities well into northern Europe: "Lombard" designated the merchant or banker coming from northern Italy (see, for instance, Lombard Street in London). The name "Lombardy" came to designate the whole of Northern Italy until the 15th century and

Figure 12: This figure shows places with connections with *London*. In this case demonstrating the globalization of northern Italy during the Renaissance.

8. REFERENCES

[1] B. Adams and K. Janowicz. On the geo-indicativeness of non-georeferenced text. In *ICWSM*, pages 375–378. The AAAI Press, 2012.

[2] B. Adams and G. McKenzie. Frankenplace: An application for similarity-based place search. In *ICWSM*, pages 616–617. The AAAI Press, 2012.

[3] G. Amati and C. J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389, 2002.

[4] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak. Spatial variation in search engine queries. In *WWW*, pages 357–366. ACM, 2008.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(4/5):993–1022, 2003.

[6] C. A. Brewer and L. Pickle. Evaluation of methods for classifying epidemiological data on choropleth maps in series. *Annals of the Association of American Geographers*, 92(4):662–681, 2002.

[7] H. Chen and S. Dumais. Bringing order to the web: Automatically categorizing search results. In *SIGCHI*, pages 145–152. ACM, 2000.

[8] Y.-Y. Chen, T. Suel, and A. Markowetz. Efficient query processing in geographic web search engines. In *SIGMOD*, pages 277–288. ACM, 2006.

[9] S. Clinchant and E. Gaussier. Information-based models for ad hoc IR. In *SIGIR*, pages 234–241. ACM, 2010.

[10] G. Cong, C. S. Jensen, and D. Wu. Efficient retrieval of the top-k most relevant spatial web objects. *Proceedings of the VLDB Endowment*, 2(1):337–348, 2009.

[11] A. S. Fotheringham and D. W. Wong. The modifiable areal unit problem in multivariate statistical analysis. *Environment and planning A*, 23(7):1025–1044, 1991.

[12] M. F. Goodchild, H. Guo, A. Annoni, L. Bian, K. de Bie, F. Campbell, M. Craglia, M. Ehlers, J. van Genderen, D. Jackson, A. J. Lewis, M. Pesaresi, G. Remetey-Fülöpp, R. Simpson, A. Skidmore, C. Wang, and P. Woodgate. Next-generation digital earth. *PNAS*, 109(28):11088–11094, 2012.

[13] K. M. Gorski, E. Hivon, A. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann. HEALPix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2):759, 2005.

[14] C. A. Gotway and L. J. Young. Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648, 2002.

[15] S. Graham and P. Healey. Relational concepts of space and place: Issues for planning theory and practice. *European Planning Studies*, 7(5):623–646, 1999.

[16] L. Gravano, V. Hatzivassiloglou, and R. Lichtenstein. Categorizing web queries according to geographical locality. In *CIKM*, pages 325–333. ACM, 2003.

[17] R. W. Gray. Exact transformation equations for Fuller's world map. *Cartographica*, 32(3):17–25, 1995.

[18] B. Gretarsson, J. O'Donovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman, and P. Smyth. Topicnets: Visual analysis of large text corpora with topic modeling. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2):23, 2012.

[19] B. Hecht, S. H. Carton, M. Quaderi, J. Schöning, M. Raubal, D. Gergle, and D. Downey. Explanatory semantic relatedness and explicit spatialization for exploratory search. In *SIGIR*, pages 415–424. ACM, 2012.

[20] B. Hecht and M. Raubal. GeoSR: Geographically explore semantic relations in world knowledge. In L. Bernard, A. Friis-Christensen, and H. Pundt, editors, *The European Information Society*, pages 95–113. Springer Berlin Heidelberg, 2008.

[21] G. F. Jenks. The data model concept in statistical mapping. *International Yearbook of Cartography*, 7(1):186–190, 1967.

[22] C. Keßler, K. Janowicz, and M. Bishr. An agenda for the next generation gazetteer: Geographic information contribution and retrieval. In *SIGSPATIAL*, pages 91–100. ACM, 2009.

[23] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[24] B. Kules, R. Capra, M. Banta, and T. Sierra. What do exploratory searchers look at in a faceted search interface? In *JCDL*, pages 313–322, New York, NY, USA, 2009. ACM.

[25] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 2014.

[26] C. D. Manning, H. Schütze, and P. Raghavan. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 2008.

[27] G. Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.

[28] D. Newman, T. Baldwin, L. Cavedon, E. Huang, S. Karimi, D. Martinez, F. Scholer, and J. Zobel. Visualizing search results and document collections using topic maps. *Web Semantics*, 8(2):169–175, 2010.

[29] L. T. Nowell, R. K. France, D. Hix, L. S. Heath, and E. A. Fox. Visualizing search results: Some alternatives to query-document similarity. In *SIGIR*, pages 67–75. ACM, 1996.

[30] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999.

[31] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR*, pages 275–281. ACM, 1998.

[32] K. Sahr, D. White, and A. J. Kimerling. Geodesic discrete global grid systems. *Cartography and Geographic Information Science*, 30(2):121–134, 2003.

[33] B. Shneiderman, D. Feldman, A. Rose, and X. F. Grau. Visualizing digital library search results with categorical and hierarchical axes. In *ACM DL*, pages 57–66. ACM, 2000.

[34] B. W. Silverman. *Density estimation: for statistics and data analysis. Monographs on Statistics and Applied Probability 26*. Chapman and Hall/CRC, 1986.

[35] D. A. Smith. Detecting and browsing events in unstructured text. In *SIGIR*, pages 73–80. ACM, 2002.

[36] J. P. Snyder. An equal-area map projection for polyhedral globes. *Cartographica*, 29(1):10–21, 1992.

[37] H. Southall, R. Mostern, and M. L. Berman. On historical gazetteers. *International Journal of Humanities and Arts Computing*, 5(2):127–145, 2011.

[38] G. R. Terrell and D. W. Scott. Variable kernel density estimation. *The Annals of Statistics*, 20(3):1236–1265, 1992.

[39] T. Tezuka, T. Kurashima, and K. Tanaka. Toward tighter integration of web search with a geographic information system. In *WWW*, pages 277–286. ACM, 2006.

[40] W. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(Supplement):234–240, Jun 1970.

[41] H. Wallach, D. Mimno, and A. McCallum. Rethinking LDA: Why priors matter. In *NIPS*, pages 1973–1981, 2009.

[42] R. W. White and R. A. Roth. Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–98, 2009.

[43] B. Wing and J. Baldridge. Simple supervised document geolocation with geodesic grids. In *ACL*, pages 955–964. ACL, 2011.

[44] Y. Zhou, X. Xie, C. Wang, Y. Gong, and W.-Y. Ma. Hybrid index structures for location-based web search. In *CIKM*, pages 155–162, New York, NY, USA, 2005. ACM.