Interactive Exploration of Geographic Regions with Web-based Keyword Distributions

Chandan Kumar
University of Oldenburg,
Oldenburg, Germany
chandan.kumar@uni-oldenburg.de

Dirk Ahlers
NTNU – Norwegian University of Science and Technology,
Trondheim, Norway
dirk.ahlers@idi.ntnu.no

Wilko Heuten
OFFIS – Institute for Information Technology,
Oldenburg, Germany
wilko.heuten@offis.de

Susanne Boll
University of Oldenburg,
Oldenburg, Germany
susanne.boll@uni-oldenburg.de

ABSTRACT

The most common and visible use of geographic information retrieval (GIR) today is the search for specific points of interest that serve an information need for places to visit. However, in some planning and decision making processes, the interest lies not in specific places, but rather in the makeup of a certain region. This may be for tourist purposes, to find a new place to live during relocation planning, or to learn more about a city in general. Geospatial Web pages contain rich spatial information content about the geo-located facilities that could characterize the atmosphere, composition, and spatial distribution of geographic regions. But the current means of Web-based GIR interfaces only support the sequential search of geo-located facilities and services individually, and limit the end users on abstracted view, analysis and comparison of urban areas. In this work we propose a system that abstracts from the places and instead generates the makeup of a region based on extracted keywords we find on the Web pages of the region. We can then use this textual fingerprint to identify and compare other suitable regions which exhibit a similar fingerprint. The developed interface allows the user to get a grid overview, but also to drill in and compare selected regions as well as adapt the list of ranked keywords.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces

Keywords

Geographic information retrieval, Spatial Web, Geographic regions, Keyword distributions, Visualization, Interaction

1. INTRODUCTION

Geospatial search has become a widely accepted search mode offered by many commercial search engines. Their interfaces can easily be used to answer relatively simple requests such as "*restaurant in Berlin*" on a point-based map interface, which additionally gives extended information about entities [1]. A corresponding strong research interested has developed in the field of geographic information retrieval, e.g., [2, 17, 15]. However, there are many tasks in which the retrieval of individual pinpointed entities such as facilities, services, businesses, or infrastructure cannot satisfy user's more complex spatial information needs.

To support more complex tasks we propose a new retrieval method based on entities. For example, sometimes the distribution of results on a map can already inform certain views about areas, e.g., a search for "bar" may show a clustering of results that can be used for "eyeballing" a region of nightlife even without sophisticated geospatial analysis. However, as users become more used to local search, more complex search types and supporting analysis are desired that enable a combined view onto the underlying data [10]. Exploration of geographic regions and their characterization was found as one of the key desire of local search users in our requirement study [11]. A person who is moving to a new area or city would like to find similar neighborhoods or regions with a similar makeup to their current home. It might not even be the concrete entities, but rather the atmosphere, composition, and spatial distribution that make up the "feeling" of a neighborhood that best capture the intention of a user. To assess this similarity of regions we propose a spatial fingerprint (query-by-spatial-example) that acts as an abstracted view onto the same point-based data.

We also aim to provide new visual tools for the exploration of geographic regions. While the necessary multi-dimensional geospatial data is already available, there is no suitable interface to query them, let alone to deal with the multi-criteria complexity. In this paper we describe a visual-interactive GIR system to support the retrieval of relevant geospatial regions and enable users to explore and interact with geospatial data. We propose a new query-by-spatial-example in-

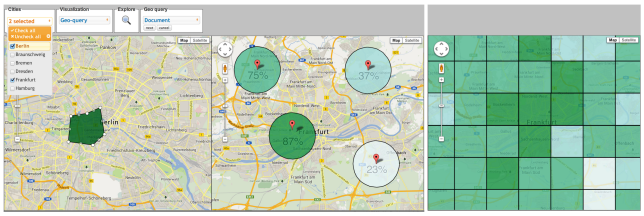

Figure 1: Geographic querying and ranking of geographic regions, with user-selected target regions and alternative grid view

teraction method in which a user-selected region's characteristic is fingerprinted to present similar regions. Users can interactively refine their query to use those characteristics of a region that are most important to them. For a more detailed overview, we use the full text of georeferenced Web pages for queries and analysis. This work goes beyond conventional GIR interfaces as it allows users to interact with aggregated spatial information via spatial queries instead of only textual querying, which is especially important to define regions of interest. We discuss the necessary input, visualization, comparison, refinement, and ranking methods in the remainder of this paper.

2. USING THE GEOSPATIAL WEB TO CHARACTERIZE GEOGRAPHIC REGIONS

The distribution of geo-entities is used to illustrate the characteristics and dynamics of a geographic region. A geo-entity is a real life entity at a physical location, e.g., a restaurant, theatre, pub, museum, business, school, etc. To open these entities up for aggregate and multi-criteria region characterization, they need a certain depth of information associated with them. It is obvious that only position information or the name of a place is insufficient, so categorial or textual description is needed. For initial studies [11, 9] we used OpenStreetMap (OSM)[1] which uses a tagging system for categories. To better characterize the geo entities we now use their associated Web pages. The reason for this is the massive increase of the amount of usable data. The Web pages of entities contain a lot more than just the basic information and can therefore be used to uncover much more detailed information. This method can also include additional sources such as events happening in the region or user-generated content on third-party pages [2]. We later describe how we identify the most meaningful keywords from the pages for this task.

To actually make the connection from a location to Web pages, we assume that the presence of location references on a page is a strong indication that the page is associated with the entity at that location. We use our geoparser to extract location references and thereby assess the geographical scopes of a page. The geoparser is trained to the presence of location references in the form of addresses within the page content. This is a suitable approach for the urban areas we are addressing in this work, because we need a geospatial granularity at the sub-neighborhood level. Knowledge-based identification and verification of the addresses is done against a gazetteer extended with street names, which we

[1] `http://www.openstreetmap.org/`

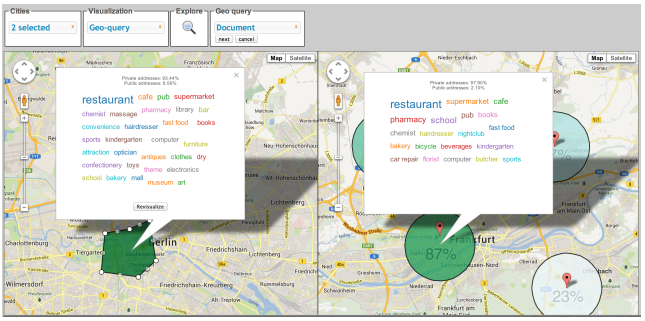

Figure 2: Keyword-based visual comparison of geographic regions

fetched from OSM for the major cities of Germany. To retrieve actual pages, we crawled the Web with a geospatially focused crawler [3] based on the geoparser and built a rich geo-index for various cities of Germany, where each city contains several thousand geotagged Web pages with their full textual content.

3. INTERFACE FOR EXPLORATION OF GEOGRAPHIC REGIONS OF INTEREST

We have implemented two main interaction modes in the Web interface as shown in Figure 1. A user intends to compare multiple geographic regions of Frankfurt (target region, right in the dual-map view) with respect to a certain relevant region in Berlin (query region, left). The current reference region of interest is specified via a visual query. The user can then either select regions by placing markers onto the map, or alternatively use a grid overview (right side of Figure 1). In both cases, the system computes the relevance of the target regions with respect to the characteristics of the query region.

3.1 Query-by-spatial-example

Most GIR interfaces use a conventional textual query as input method to describe user's information need or use the currently selected map viewport. We wanted to give users the ability to arbitrarily define their own spatial region of interest. The free definition of the query region is important, as users may not always want a neighborhood that is easily describable by a textual query. We therefore enabled to query by spatial example, where users can define the query region by drawing on map. Figure 1 shows an example of a user selected region of interest via a polygon query (by mouse clicks and drag) in the city of Berlin.

3.2 Visualization of suitable geographic regions

Users can select several location preferences in their target region that they would like to explore by positioning markers on the map interface. The system defines the targets with a circle around the user-selected locations with the same diameter as the reference region polygon. The target regions obtain the ranking with respect to their similarity with the reference region. Their relevance is shown by the percentage similarity and the heatmap based relevance visualization. We used a color scheme of different green tones which differed in their transparency. Light colors represented low

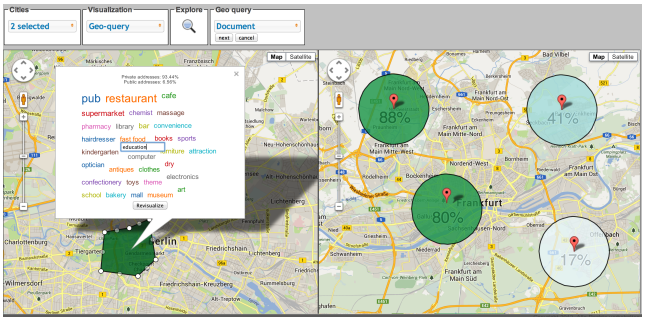

Figure 3: User interaction with the keyword distribution and revisualization

relevance, dark colors were used to indicate high relevance. The color scheme selection was aided by ColorBrewer [2].

As an example, Figure 1 shows 4 user-selected locations on the city map of Frankfurt, the circle regions around these 4 markers have the same diameter as the query region in Berlin. The target region in the centre of the city is most relevant with the similarity of 88%, and consequently has the darkest green tone. If a user has not yet formed any preference, we offer an aggregate overview of geo-entities. We partition the map area using a grid raster [14], as we do not intend to restrict user exploration to only selected areas. There could be situations when users look beyond the specific target regions, and would like to have an overview of the whole city with respect to a query region. The right side of Figure 1 shows the aggregated ranked view of the grid-based visualization. Each grid cell represents the overall relevance with respect to the query region. The visualization gives a good overview and assessment on relevant regions which the user can then explore further. Users can select the grid size, which is otherwise similar to the size of the query region. The grid layout is fixed to the city boundaries as we intend to give the overview of whole city. In the future we would like to make it more dynamic where users should be able to shift the grid layout, since a slight variation in grid cell boundaries could alter the relevance results.

3.3 Exploration and interaction with geographic regions via keyword distributions

Interaction models should provide end users the opportunity to explore the characteristics of selected regions, and adapt it further to their requirements. We initially show the most relevant keywords of the respective region using a word cloud. The word cloud provides more detailed information on keyword distribution when the mouse hovers over it. The font size and order of the keywords signify their relevance. Figure 2 shows the comparison of the query region with the most relevant target region via both their keyword distributions. In this case, the distributions of both regions are very similar, leading to the high relevance score for the target region.

Since the keyword characteristics of a query region is derived from the georeferenced Web pages, there are situations where a user might not be satisfied with the spatial description and wants to influence the keywords. In the example of Figure 3, a user decides that pubs are more important than restaurant, fast food is not an aspect of his lifestyle and should be replaced by education facilities near his new home. In such scenarios users need to interact and adapt the generated keyword distributions of query regions. We make the word cloud interactive and editable. Users can drag keywords to alter their position and thus their significance. They can also edit, delete or replace keywords in the word cloud to change the criteria. After modifying the keyword distribution, users can revisualize the target regions to update their ranking. Figure 3 shows this user interaction with the word cloud, including the revisualization of the updated ranking of target regions, which are visibly different from the previous ranking of Figure 2.

[2] `http://colorbrewer2.org`

4. TEXT-BASED CHARACTERIZATION AND RANKING OF GEOGRAPHIC REGIONS

We adapt common IR methods for ranking and similarity measures. In relevance-based language models, the similarity of a document to a query is the probability that a given document would generate the query [12]. To be able to do the same with geographic regions, we add a transitional step. Regions are considered as compound documents built from the Web pages of the entities inside them. We can then define the similarity of document clusters of regions based on the probability that the target region can generate the query region. The Kullback-Leibler divergence is used for comparison [4].

For a geospatial document d, we estimate $P(w|d)$, which is a unigram language model , with the maximum likelihood estimator, simply given by relative counts: $P(w|d) = \frac{tf(w,d)}{|d|}$, here $tf(w,d)$ is the frequency of word w in the document d and $|d|$ is the length of the document d. A geographic region contains several geospatial documents insides its footprint area. We define a geographic region based on a document cluster D which contains document $\{d_1, d_2....d_k\}$, and the distribution of a particular word w in the geographic region would be estimated with its combine probability in the collection $P(w|D) = \frac{1}{k}\sum_{i=1}^{k} P(w|d_i)$. The word cloud represents the most prominent keywords of the region with respect to their ranked probability distribution $P(w|D)$. The comparison of regions is done with respect to their probability distribution using KL-divergence. A target region x will be compared to the query region as following

$$Relevance(Region_x) = \sum_w P(w|D_q) log \frac{P(w|D_q)}{P(w|D_x)}$$

The computation of this formula involves a sum over all the words that have a non-zero probability according to $P(w|D_q)$. Each region $Region_x$ gets a relevance score according to its distribution comparison to the query region $Region_q$. All target regions (user selected regions or grid based divisions) are ranked with respect to their relevance score for visualization.

5. RELATED WORK

The field of geographic information retrieval examines documents' geospatial features at a regional scale and also at smaller granularities and usually supports keyword@location queries [2, 17, 15]. Similarly, location-based services (e.g.,

FourSquare, Yelp, Google Maps) allow users to retrieve and visualize geo-entities matching a category or search term. However, search for multiple categories or other complex tasks is usually not supported. Some non-conventional spatial querying methods have been proposed, e.g., query-by-sketch on a map [6]. Other work uses the density of arbitrary user-supplied keywords to build a query region [8]. Tag clouds have been adapted to maps, exploiting georeferenced tags [16]. Locally characteristic keywords can be extracted for map visualization and to show their spatial extent [19]. None of these approaches make a larger word cloud available, but only the main terms. Other geovisualization approaches [5, 7] approach multi-criteria analysis, but are usually targeted to specific domains and experts. The Inspect system was tailored at geospatial analysts to visually filter and explore multidimensional data [13]. A multi-criteria evaluation for home buyers was proposed in [18]. The scenario of spatial decision making is similar to ours, but it focused on experts and spatial computation issues rather than interface and visualization aspects.

Our system interface differs in the granularity of information need and representation, i.e., we focus on the ranking of regions, but base it on high-granularity geo-entities that have a very exact location, which ensures that the spatial query does not produce overlap to neighboring regions and makes the multi-criteria analysis more exact to be executed at arbitrary region sizes.

6. CONCLUSIONS AND FUTURE WORK

Most current local search interfaces do not offer adequate support for the exploration and comparison of geographic areas and regions. End users need visual and interactive assistance from GIR systems for an abstracted overview and analysis of geospatial data. We proposed interactive interfaces for the characterization and assessment of relevant geographic regions that enable end-users to query, analyze and interact with the rich geospatial data available on the Web in user-selected geographic regions. The relevance of regions is based on the similarity of keyword distributions.

The observation of results shows satisfactory performance by uncovering realistic and meaningful keywords defining the regions. We observed that the characterization and comparison of geographic regions show good results with respect to geo-located facilities and infrastructure of German cities, e.g., clearly distinct characteristics for university, industrial, or party districts. In the future we plan a more formal qualitative and quantitative evaluation of these interfaces, to examine the acceptance of these visualizations with regard to user-centered aspects such as exploration ability, information overload, and cognitive demand. We would also like to explore more advanced interaction methods to enhance the usability of the proposed visualizations.

Additionally, we envision more powerful region similarity measures such as landscape and topological similarity, similarity via social media, and an integration of additional data sources.

Acknowledgments

The authors are grateful to the DFG SPP 1335 'Scalable Visual Analytics' priority program which funds the project UrbanExplorer. The 2nd author acknowledges funding from the ERCIM "Alain Bensoussan" Fellowship Programme.

7. REFERENCES

[1] D. Ahlers. Local Web Search Examined. In *Web Search Engine Research*. Emerald, 2012.

[2] D. Ahlers and S. Boll. Location-based Web search. In *The Geospatial Web*. Springer, 2007.

[3] D. Ahlers and S. Boll. Adaptive Geospatially Focused Crawling. In *CIKM '09*, 2009.

[4] T. M. Cover and J. A. Thomas. Elements of information theory, 1991.

[5] J. Dykes, A. M. MacEachren, and M.-J. Kraak. *Exploring Geovisualization*. Elsevier, 2005.

[6] M. J. Egenhofer. Query processing in spatial-query-by-sketch. *J. Vis. Lang. Comput.*, 8, 1997.

[7] R. Greene et al. GIS-based multiple-criteria decision analysis. *Geography Compass*, 5(6), 2011.

[8] A. Henrich and V. Lüdecke. Measuring Similarity of Geographic Regions for Geographic Information Retrieval. In *ECIR '09*, 2009.

[9] C. Kumar, W. Heuten, and S. Boll. Visual interfaces to support spatial decision making in geographic information retrieval. In *CD-ARES 2013*. to appear.

[10] C. Kumar, W. Heuten, and S. Boll. Geovisualization for end user decision support: Easy and effective exploration of urban areas. In *GeoViz_Hamburg 2013: Interactive Maps That Help People Think*, 2013.

[11] C. Kumar, B. Poppinga, D. Haeuser, W. Heuten, and S. Boll. Geovisual interfaces to find suitable urban regions for citizens: A user-centered requirement study. In *UbiComp'13 Adjunct*, 2013. to appear.

[12] V. Lavrenko and W. B. Croft. Relevance based language models. SIGIR '01. ACM, 2001.

[13] S.-J. Lee et al. Inspect: a dynamic visual query system for geospatial information exploration. In *SPIE*, 2003.

[14] A. M. MacEachren and D. DiBiase. Animated maps of aggregate data: Conceptual and practical problems. *CaGIS*, 18(4), 1991.

[15] A. Markowetz et al. Design and Implementation of a Geographic Search Engine. In *WebDB 2005*, 2005.

[16] D.-Q. Nguyen and H. Schumann. Taggram: Exploring geo-data on maps through a tag cloud-based visualization. In *IV'10*, 2010.

[17] R. S. Purves et al. The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the internet. *IJGIS*, 21(7), 2007.

[18] C. Rinner and A. Heppleston. The spatial dimensions of multi-criteria evaluation – case study of a home buyer's spatial decision support system. In *Geographic Information Science*. 2006.

[19] B. Thomee and A. Rae. Uncovering locally characterizing regions within geotagged data. WWW '13, 2013.