Metro-Wordle: An Interactive Visualization for Urban Text Distributions Based on Wordle

Chenlu Li*a, Xiaoju Donga,∗, Xiaoru Yub

aDepartment of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, P.R. China, 200240

bKey Laboratory of Machine Perception (Ministry of Education), and School of EECS, Peking University, Beijing, P.R. China, 100871

ARTICLE INFO

Article history:
Received 11 December 2017
Received in final form 23 February 2018
Accepted 12 March 2018

Keywords: Text visualization, Location retrieval, Urban data, Metro map, Word cloud

ABSTRACT

With the development of cities and the explosion of information, vast amounts of geo-tagged textural data about Points of Interests (POIs) have been generated. Extracting useful information and discovering text spatial distributions from the data are challenging and meaningful. Also, the huge numbers of POIs in modern cities make it important to have efficient approaches to retrieve and choose a destination. This paper provides a visual design combining metro map and wordles to meet the needs. In this visualization, metro lines serve as the divider lines splitting the city into several subareas and the boundaries to constrain wordles within each subarea. The wordles are generated from keywords extracted from the text about POIs (including reviews, descriptions, etc.) and embedded into the subareas based on their geographical locations. By generating intuitive results and providing an interactive visualization to support exploring text distribution patterns, our strategy can guide the users to explore urban spatial characteristics and retrieve a location efficiently. Finally, we implement a visual analysis of the restaurants data in Shanghai, China as a case study to evaluate our strategy.

© 2018 Published by Elsevier B.V. on behalf of Zhejiang University and Zhejiang University Press. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Rapid urban growth enriches people’s life and gives people more choices about where to go, but it also becomes more difficult for people to select one from numerous POIs. At present, many websites provide discussions, reviews and descriptions on POIs to help users make a choice, therefore a large number of textual data has been generated. However, people can be really lost in confront of the huge data. It is in great need efficiently finding the wanted location. For the text data on POIs has a close connection with locations, it is meaningful to visualize the texts combining with their geographic information. In addition, metro has been a popular urban transportation choice, and subway lines are like the skeleton of a city, so we use the metro lines to separate urban space into several areas, then we study how the texts distributed across the areas and try to find their spatial characteristics.

Traditional methods to observe and retrieve locations include maps and search engines. As for maps, designers have made many kinds of colorful maps for tourists to explore the city, such as chinamaps.org and vidiani.com. These maps are well designed and attractive, but are mostly static with a small amount of information and take a lot of manpower. Some online web mapping services like Google Map support interactive searching, however, they are lack of detailed descriptions. As for search engines, the results obtained by queries are usually listed as text documents which are short of geo information. It’s difficult to compare a large number of POIs using the textual and geo information at the same time through the existed methods. At present, there is much research aiming at analyzing geo-tagged data of POIs. Lee et al. (2011), Yin et al. (2011) and Xu et al. (2017) analyze text data combining their locations to discover hidden spatial patterns, however, in these works, the text view and map view are arranged side by side which adheres to observe their relations directly. There are also some
works aiming to generate a more intuitive overview of how semantic information distribute across space. Wu et al. (2013), Afzal et al. (2012), BRA and Meitzler are some examples of visualizations that merge text and spatial data into one graph. They considered aesthetics and user-friendliness, but they are more suitable for displaying rather than exploring data because of their small amount of information and insufficient interactions.

As a result, we want to design a visualization that achieves the following objects:

- All POIs are accessible through keywords. In the main view, we only provide a certain number of keywords due to the space limitation, but via these words and with the help of the filters, users can retrieve any POIs in the city.
- Textural and geographical information can be observed in one visual representation. So that their spatial relations and characteristics can be discovered.
- Multiple dimensions data of POIs are presented to give users comprehensive information.

To achieve the goals we proposed above, this paper introduces a visual approach to integrate text and metro map. First, we used the semantic analysis of POIs records and obtained some keywords after extracting and clustering the text documents. Then, we automatically calculate the layout of metro map based on Mixed-Integer Programming (MIP) algorithm. After getting the keywords of each area, we visualized the words as wordles and embedded them into the metro map using a greedy algorithm. Finally, we design a user interface to support rich interactions. We use the metro-wordle graph as the main view with several other visual models and a variety of operations for users to explore in the system.

The main contributions of this article are as follows:

- We introduce and implement a new visual strategy to combine text and map. We use the metro lines as the boundaries for wordles, and in each region, we place each word considering its actual geographic location. This method supports efficiently discovering the pattern of text spatial distributions.
- We propose new requirements for automatic drawing metro map. We take the amount of information in each region into consideration when schematizing a metro map.
- We present an interactive visualization to support urban geo-tagged text analysis. This work helps people retrieve and select POIs from large amount data efficiently. It also allows users to explore the patterns of urban topic spatial distributions.

The sections are organized as follows. This paper discusses related works on the automatic layout technique for metro maps, wordle embedding, textual and urban data visualization in Section 2. Section 3 gives an overview of this work. Section 4 explains the methods of data analysis including words extraction methods, metro map layout algorithm and wordle embedding. Section 5 presents the visual design of the system. Then we use the metro-wordle map for restaurants in Shanghai as a case study in Section 6. In the end, we discuss and conclude our work.

2. Related Work

2.1. Wordle

Numerous text-based visualization methods have been proposed. We choose wordle as the main method to visualize text in our work. Wordle was launched in 2009 (Viégas et al. 2009).
Visual Analysis of Urban Data

Development of city has generated huge data including the geographic information, textual data, transportation data and so on. A variety of methods have been proposed to visualize urban data in different dimensions.

Geographic location is an important attribute of urban data because many dimensions need to be combined with it to see how they distribute in space. However, the classic statistical analysis and geographic information system (GIS) are difficult to exploit the complex relations in multiple dimensions. So visual analysis technique can be very helpful in this situation to guide the users to explore the geospatial data. As for textual data, people like to air their opinion on open web platform like social networks and recommend systems. Excavating the hidden semantic information in texts is meaningful. Oelke et al. (2009) analyzed customer comments and ratings to determine the positive and negative opinions and visualized the results of opinion analysis. Wu et al. (2010) designed an interactive visualization system to support complex opinion analysis of online hotel customer feedback. As for transportation data, it includes data of public transit facilities and data generated by human activities such as trajectory data from which we can study the pattern of urban traveling behavior to help improve urban transportation services.

In many cases, we want to observe and explore the different dimensions of data in one system. Each dimension implies some important features of the city in a different perspective and is related to other dimensions at the same time. Studying them together can give us a more comprehensive understanding of the city. Lee et al. (2011) presented a social event detector for finding out unusually crowded places based on the geo-social event on social networking sites. Yin et al. (2011) introduced a model integrating text and maps to extract topics from Flickr database with photos tagged by locations to analyze the geographical characteristics of different topics. Xu et al. (2017) proposed a system exploring spatiotemporal urban topics using reviews data. These works have merged text data and their locations to discover hidden spatial patterns, however, in these papers text analysis module and geographic module were arranged side by side and lacking visual relations. There are also works generating a more straightforward visualization to observe how semantic information distributes across space. Wu et al. (2013) presented an approach for enhancing metro maps by placing large annotation labels around metro stations. Afzal et al. (2012) introduced a method for generating typographic maps automatically where words form the graphical features. BRA, Meitzler are also examples of visualizations that merge text and spatial data into one view. However, they were not suitable for exploring large data because of insufficient interactions.
To solve existing defects mentioned above, our work focuses on combining text and metro map in one visual representation and provide ability to interactively explore large geo-tagged textual data.

3. Overview

To achieve our aims, we need to develop four tasks of analysis.

- **T1: Obtain the keywords of POIs.** Words in wordles are extracted from the textual data about POIs. As we can’t place all of the relevant words in the wordle due to space limitation, we need to get an appropriate amount of keywords. These keywords should not only cover all POIs but also attract users. Besides the descriptions and the comments on POIs, we can include other information in keywords. For example, if the topic is food and we want to retrieve restaurants (the POIs), we can choose the signature dishes of the restaurants as the keywords; if the topic is sightseeing and we want to look for sight spots to visit, we can choose the attractions as the keywords. The keywords will be clustered to fit in the region.

- **T2: Abstract the metro map and divide the urban space into several subareas.** What we concerned more about the metro map is the topological relations of all stations. So we need to abstract the topological structure from metro map. Then we calculate an aesthetic layout of the metro map based on design criterion. For we analyze the urban characteristics and text distributions mainly on a subarea level, we need to figure out all the sub-regions in the metro map.

- **T3: Place words into each subarea and generate constrained wordles.** Wordles are constrained by the metro lines. The words we get all come with geographical data. The positions of the words on a map should be relative to their actual geographical locations, at the same time, to some extent the positions remain flexible to adaptively fit with the sizes and shapes of the regions where they stay.

- **T4: Provide an interactive visualization.** Our goal is to get an interactive visualization which supports data exploration rather than a static map. Besides a metro-wordle map as an overview to observe the overall distributions and features, we also need to provide some visual modules for filtering and details.

To achieve these tasks, our visualization consists of data analysis and visual design. Fig. 2 shows the pipeline of our system. The input of the system is the data set, containing the POIs records and the metro information. The data analysis module first collected the textual data from the Internet, then the keywords are extracted and clustered. Then we abstract metro map as a topological graph and figure out an optimal layout. Next, the keywords are put into subareas and generate wordles. The details of the data analysis are described in Section 4.

As for the visual design, three main interactive views are designed, namely, Metro-Wordle View, Geographic View, Statistics View, and Filter View. Metro-Wordle view shows the embedding wordles in a metro map. Geographic View shows the actual geo locations of POI, Statistics and Filter View shows the statistics data of POI and provides filters. These views are linked to each other. The details of the visual design are described in Section 5.

4. Data Analysis

4.1. Keywords Extraction

We collected POIs information on a particular topic from the Internet. We mainly used the geo-tagged textual data for our visualization. However, the names of POIs are not suitable to be placed on map directly as their numbers are too large to be put into the limited space and the abandonment of the POIs will cause information loss. So we used the descriptive words for POIs to generate wordles. Relatively small number of the descriptive words can cover all POIs.

We extracted descriptive words from users’ comments on POIs or other textual properties of POIs (e.g., tags). After getting words, we measured their similarities to merge word items and filtered the useless words. As the number of words were limited by the area of region, we clustered the similar words to control their numbers. The similarities between words are measured based on their geographical locations and semantic similarities, only the words that are both semantically similar and in close positions can be clustered by k-means. The parameters $k_j$ for clustering are related to their original number and the space of area. The clustering parameters were optimized along with the size of a region using MIP method introduced in Section 4.4.

Then we gave each cluster a single word and it was assigned a weight based on the number of words in the cluster.

4.2. Abstract the Metro Map

As what we concern about a metro network mainly is its topological structure, we abstract the metro map as a topological graph. First, we extracted the ”key nodes” of the metro map. We defined the metro stations satisfying one of following conditions as ”key nodes”:

- The station is a departure station or terminal station.
- The station is a transfer station.
- The station connects two lines between which the angle is not within (0, 45°) nor (135°, 180°).

We ignore other nodes except ”key nodes” and transform the metro network into a graph $(G)$. The ”key nodes” of the network correspond to the set of vertices $(V)$, the links between pairs of stations correspond to the set of edges $(E)$. So the metro layout problem is modeled as a graph drawing problem, which aims to find a suitable geometric representation of a graph $G = (V, E)$. This problem transformation simplifies the network and provides a basis for the following layout optimization process.
After getting the optimized layout of G, the ignored stations would be inserted into the line segments where they stayed according to their distances from the key nodes.

Edges divided the plane into a set of regions. As for the unbounded region, we extended the line segments and used the boundaries of the canvas to bound them.

4.3. Statistical Analysis for Subareas

Keywords have been tagged with locations (their related POIs’ latitude and longitude). The map has been divided into several closed sub-regions by metro lines both in the actual geographic space and the schematized space. Using the geographic locations of a point and the boundaries of the region, we figured out whether a point is in or out the region. After figuring out to which subarea each word belongs in actual geographic space, we recorded the relative location \((dx, dy)\) of each word and the region it stays. Based on the relative positions, the points and words would be put into the distorted metro map and stay in the same sub regions in the schematized metro map and the actual geographic metro map. We calculated the area for each region denoted by \(A_{2j}\) for the \(j\) – \(th\) region. We also collected other statistical data about POIs (e.g. grades, prices, popularities) to observe their geographic distributions and the each region’s characteristics in multiple dimensions.

4.4. Automatic Layout of Metro Map

The metro map should be distorted to meet the following rules:

- R1. All line segments are restricted to the four octilinear orientations: horizontal, vertical, and 45-diagonal.
- R2. The geographical network topology and the relative position between stations are remained to support the mental map of the users.
- R3. The bends along individual metro lines should be avoided. If bends cannot be avoided, obtuse angles are preferred over acute angles.
- R4. The length of lines segments between two vertices of the abstract graph should be kept as uniform as possible.
- R5. The size of subareas should be balanced with the number of POIs in it.

The Rule 1-4 were borrowed from basic design rules to which almost all schematic metro maps conform and were introduced by K. Garland (1994) in the first modern metro map, and we also made reference to the layout principles of official metro maps (e.g. Ovenden (2003), Roberts (2005), Afzal et al. (2012)). Rule 1-4 are to achieve the topological consistence and aesthetics. We added the R5 to make the area of each region fit into its text information. These rules were set to find a
compromised layout between the geographical locations, user perceptions and the amount of textual information.

To follow above rules, we used the MIP algorithms for we could easily add our constraints and change the target function to meet our own needs.

According to the rules, our hard constraints are:

- **H1.** For each edge, the line segment must be octilinear.
- **H2.** For each vertex, the circular order of its neighbors must remain as the input.
- **H3.** For each edge, the line segment must have a minimum length.
- **H4.** Each edge must have distance from each non-incident edge to avoid the intersection.

The hard constraints are set to achieve R1-R2 and ensure the readability of the metro map.

The soft constraints are as follows:

- **S1.** The lines should have few bends and the bend angles (<180°) should be as large as possible.
- **S2.** For each pair of adjacent vertices, their relative position should be preserved.
- **S3.** The total edge length should be small.
- **S4.** For each area, the ratio of the sum of the weights of all words in it to the size of its space is as close to 1 (the optimal ratiom from experiments) as possible.

The soft constraints are set to achieve R2-R5.

To apply MIP method, hard constraints are modeled as sets of equations and inequations, and soft constraints are modeled as a cost function. H1-H4 and S1-S3 were borrowed from [Nöllenburg and Wolff (2011)]. You could refer to paper [Nöllenburg and Wolff (2011)] to find the detailed formulas. We added the amount of information into consideration when generating metro layout. So we discussed the S4 in detail as it was novel.

We assume that the size of a character is a constant $S$, the clustering parameter in $i$th subarea is $k_i$, the number of letters in $i$th word is $n_i$, the space occupied by all words in $i$th subarea is $A_1 = \sum(n_i * S / k_i)$, the area of $j$th region is $A_2j$. In each region when $A_1$ is close to $A_2$, the region fits the words in it well. We add the formula $\sum_j(1 - (A_1j/A_2j))$ into the MIP’s cost function to achieve this requirement.

As we modeled our questions as a MIP problem, our aim is to minimize the cost subjected to the constrains. It is an iterative procedure to find an equilibrium configuration for this MIP problem. The parameters we want to optimize are the locations of the metro stations in schematized map $(x_h, y_h)$ and the clustering parameters for each sub region $k_j$. After getting all constraints modeled as equations and cost functions, we use CPLEX to help us solve it and get an optimal solution. We obtained and recorded all the parameters in the optimal solution.

4.5. Wordle Embedding

After getting the keywords and the map, the final step is to place the words into the metro map. A classic method of wordle embedding is using a collision detection and greedy algorithm (e.g. Koh et al. (2010), Strobelt et al. (2012)). Our method is based on it, in addition, we take the geographic locations of words into consideration. Buchin et al. (2016) also generated wordles considering their geo locations. However, their methods calculated the errors between the space that original points occupied and the area of words to minimize the errors, and took too much time to be interacted with online. In our method, the number of words, the occupied space of words, and clustering parameters were all constrained in MIP model, so we did not need to calculate the spacing errors as [Buchin et al. (2016)], which can save much time.

We first calculated the locations of the words on the distorted map based on their relative positions in actual geographic region (obtained in Section 4.3). Then when placing a word, we used its mapped location as the starting point and detected whether there was collisions with other words or subway lines. If there was a collision, we try other positions by moving along with a line shaped as the boundary of the region and around the start point until it finds an available position or if it goes too far (>5* font_size) then we would scale down the font size. All words are placed one by one according to their weights order. For each region, the sum of the area of words is corresponding with the region’s area, so we can get a compact wordle finally.

5. Visual Design

In response to our aims, we derive these requirements to guide our visual design process: users can see how the keywords distributed across the space; users can retrieve all POIs through the keywords easily; users can connect the POIs on a distorted metro map with their real-world geographic locations. We develop a visualization which integrates Metro-Wordle View, Geographic View, Statistics and Filter View, POIs Detail View. These views are linked. In the following subsections, we discuss these views and the interactions.

5.1. Metro-Wordle View

![Fig. 3. Metro-wordle view](image)

The metro-wordle view serves as the main view. It provides an overview of spatial distributions of the text. In this view, the
metro map has been distorted obeying the rules we presented in Section 4.4 and the words in it illustrate the characteristics of this area. From this view, a user can get an intuitive understanding of what each subarea have. The size of a word is proportional to its weight which is measured in Section 4.1. The color of a word presents its category. Users can pick a word from the wordle to explore the POIs related to it.

5.2. Geographic View

The geographic view presents the original locations of metro stations and POIs. From this view, users can observe how POIs distribute spatially in actual geographic space. In Fig. 4, a point represents a POI, we project it into a map based on its latitude and longitude. The color of a point denotes its category. Users can select one point and see the POI's detail information. When a point is hovered or selected, it will be large and highlighted by stroking. The nearby stations and the subway lines will also be highlighted and listed.

5.3. Statistics and Filter Module

The filter module is composed of a group of buttons and a set of charts. In Fig. 5, we get 18 categories. The buttons have corresponding category name on it, and the category will be selected once it is clicked. The group of charts show the statistic data from six aspects of POIs including average spend, popularity (measured by comment number) and scores (average, taste, environment, service). The y-axis represents the number of POIs. Users can cross-filter the data by brushing on the chart, as shown in Fig. 6. All of the views are connected and sharing the data, so when the filter changes, the views will change accordingly.

5.4. POIs Detail View

POIs detail view allows users to see the details of a POI. Metro-wordle view displays the keywords of POIs, but sometimes what we concern about is a specific place. So users should be allowed to retrieve all places through keywords in the wordle. When a word is clicked, the POIs related to it will show on the map. As shown in Fig. 7, the highlighted keyword is the one we picked and the points represent all the POIs related to the word. To see the information of a specific POI, we can click on a point, the name of it will show in a tool-tip and detailed info will show on the left side.

5.5. Interactions

Our visualization supports abundant interactions for users to explore and find hidden information. We design the visualization according to the framework: "Overview first, zoom and filter, then details-on-demand."

1. The Metro-Wordle View and the Geographic View provide the overview of the text distributions across the metro map. Users can switch between two views or arrange them side by side. During the view switching, the focus of users won’t lose.
2. Users can filter the data using the statistics and filter module or select regions in the map to focus on their interested part.
3. When a specific word is selected, the related POIs will show and be highlighted. When a POI is selected, its detailed information then appears.

6. Case Study

Shanghai is an open and hi-tech modern city. People living there are used to searching and commenting places on the Internet, which generate a huge number of data and causes the needs to retrieve and choosing restaurants. In addition, the metro system of Shanghai is mature and well designed, so the relations
between POIs and subway lines are worth studying. As a result, we choose food information in Shanghai as the input data to demonstrate the effect of our work. We query "food" on Dianping.com (a Chinese urban life guide website and a platform for sharing comments on shops). We get 9549 records of restaurants. Each record contains the shop’s detailed information, including the shop ID, geographic location (latitudes and longitudes), scores, comments, etc. Records have some properties whose values are words, such as "tag", "specialties", "category". We used these words directly. As for long descriptions and comments, keywords were obtained by sentence segmentation. We use the Python to code our MIP model, the optimization procedure runs on average in 1 minute when using four metro lines and 2 hours when using all 14 lines. We use JavaScript to draw the wordles and implement the online user interface.

6.1. Location Retrieve

We illustrate how our visual analysis system helps users find and choose a POI. Suppose that a user wants to go to eat beefsteak for dinner.

As shown in Fig. 8, regions A, B, C are three regions that contain beefsteak most obviously. Area A is enclosed by the line 1, line 4 and line 9. Area B is enclosed by the line 4, line 8 and line 9. Area C is enclosed by the line 1, line 4 and line 8. Suppose the user lives in place D, which is near the line 1 and makes it more convenient to go to region C, so he decides to go region C for dinner. Click on the words in region C, then observe the filter window. The related POIs are presented as yellow lines on the charts to denote how these candidate POIs distribute in price, taste, popularity and other dimensions. It appears that restaurants around word3 are more expensive with good quality, and the restaurants related to word2 have relatively lower scores. Restaurants around word1 have high scores and suitable price. So we choose word1 for next exploration.

Hover on the points around word1 to see details of each restaurant, compare them then we can find the best one to go.

6.2. Urban Characteristics Study

We demonstrate how our system can effectively discover the patterns of text distributions across the space.

From the wordle view, the different kinds of food have different spatial patterns. For example, there are many western-style dishes, coffee and desserts in the city center. We can speculate that it is because in this area there are many foreign companies and the major consumers are white collars and foreign visitors.

We switch to the Geographic View. In Fig. 9, we can see that the city center has much higher density of restaurants. However, some suburban regions far from the city center also have clustered restaurants, like region A, B, C. We can speculate that these regions are residential areas. Switch to the metro-wordle view, what we find interesting is that hot pots and Sichuan dishes are popular in these areas. The prices there are lower than the city center, so they are a good place to go for who want to eat pickled fish, Bullfrog, hot pot, lobster etc.

From the statistics data view, we can see that the average price per person in Shanghai is concentrated near 100 yuan. Brush the filters, the spatial distributions will vary. For example, we select the service score in range (9,10), the points on
the map become more concentrated in the city center, and the average score moves higher. We can guess that the restaurants with better service usually charge higher and are located in city center.

7. Discussions and Conclusion

In this paper, we have proposed a visualization to discover the patterns of text distributions and retrieve POIs in urban. And we have introduced a method to embed words into a metro map with the metro lines serve as constraints of wordles and words are placed on the map based on its actual geographic locations.

Semantic and statistical analysis are used to process the text. MIP method is used to draw metro map automatically. Our visualization is composed of geographic view, metro-wordle view, filters and detail information view. It also allows users to interact for further exploration. The case study in Section 6 demonstrates the effectiveness of our methods. Besides the Shanghai food map case, we can also apply our methods to study other topics. For example, if we input the data of sight points and metro information of Beijing, we can obtain an interactive Beijing traveling map. Our method can also apply to other languages. The width of a single letter and the number of the letters in word are used when calculating the area occupied by all words. When using different languages, just change the base width of letter and count the letters. We only used four metro lines in the case study as the pictures in the paper are small and we want to make sure the words can be recognized in the picture. We also add an English version (Fig. 10) and a complete version (Fig. 11) of the Shanghai metro-wordle map to show the generality of our work.

Although useful and impressive, our work still has many to improve. As for the interaction design, it can be more flexible by allowing users to tune some parameters on the interface, such as the size of one region, the number of subway lines etc. Moreover, when the positions of nodes on subway lines are adjusted, the wordles can adjust themselves correspondingly. Users should also be able to delete words that they do not interest in to get a more personalized map. In addition, different partitionings (e.g. highways, railways, rivers) can be used in the visualization. Using different dividers can get various maps which imply different spatial characteristics of text. Our method takes the dividers as several straight line segments and some properties of polygons are used, so other dividers should be straightened and regularized (e.g. extract and connect the key nodes) before applying our approaches. For our work can be developed and integrated continuously, it has potential to improve and expand in the future.

Acknowledgments

We thank the anonymous reviewers for their valuable comments. This work is supported by National Key Research and Development Program of China (Grant No. 2017YFB0701900, 2016QY02D0304) and National Nature Science Foundation of China(Grant No. 61572318, 61672055)

References


Fig. 10. English version of metro-wordle with 14 metro lines.

Fig. 11. Chinese version of metro-wordle with 14 metro lines.


