A Technique to Reduce Overlapping Symbols on Proportional Symbol Map of Multi-dimensional Data

Nadine Nandanari

(Master's Program in Computer Science)

Advised by Kazuo Misue

Submitted to Graduate School of
Systems and Information Engineering
in Partial Fulfillment of the Requirements
for the Degree of Master of Computer Science
at the
University of Tsukuba

March 2016

Abstract

Visualization of multi-dimensional data is challenging since the visualization has to able to display the multiple data dimensions as well as to maintain the comprehensible display of it. One method to visualize multi-dimensional data that has geographical location and size attribute is by using proportional symbol map, since it has symbols associated with point locations on the map and the symbols change size according to the value they represent. However, in proportional symbol map, there is a high possibility of overlapping symbols, since the size of symbols don't always proportional to the location's size they are located. Therefore, the purpose of the research is to reduce those overlapping symbols while maintaining the proportional ratio and user understanding of the data. In the proposed method of this research, overlapping symbols are merged according to the similarity of the attributes they represent and the overlapping rate of the symbols' size. Similarity is evaluated by using cosine measurement and overlapping rate is evaluated by calculating the ratio between the overlapping area of two overlapping symbols and the area outside that overlapping area. Some algorithms to execute the proposed method are developed and threshold of the similarity and overlapping area are decided by running some programs that implement the algorithms. The algorithms are evaluated by comparing some viewpoints after the running of the program, and the best algorithm is decided based on the effectiveness in reducing the overlapping symbols and the maintenance of user understanding.

Keywords: multi-dimensional data, proportional symbol map, overlapping symbols, similarity, overlapping rate, cosine measure

Contents

Abstract i
Contents ii
List of Figures iii
List of Tables iv
Chapter 1 Introduction 1
1.1. Background 1
1.2. Research Purpose 1
1.3. Structure of the Thesis 1
Chapter 2 Related Research 3
2.1. Visualization of Multi-dimensional Data 3
2.2. Usage of Proportional Symbol Map 3
2.3. Handling Overlapping Symbol 4
2.4. Clustering Based on Similarity 5
Chapter 3 Requirement and Proposed Method 6
3.1. Requirement 6
3.2. Proposed Method 6
3.2.1. Similarity Measurement 6
3.2.2. Overlapping Rate 7
3.2.3. New Value of Merged Symbols 8
3.2.4. Algorithm Development 9
Chapter 4 Use Case Data and Experiment 13
4.1. Use Case Data 13
4.2. Pre-Experiment Activity 13
4.3. Experiment 15
4.4. Analysis 22
Chapter 5 Conclusion 24
5.1. Summary 24
5.2. Future Work 24
Acknowledgement 25
Reference 26
Appendices 28

List of Figures

Figure 1. Illustration of Two Overlapped Circles 7

Figure 2. Illustration to Find Similarity of Overlapped Symbols 9

Figure 3. Illustration of Implementation 1 10

Figure 4. Illustration of Implementation 2 11

Figure 5. Visualization of original data of Race and Ethnicity in the United States on zoom level 4 13

Figure 6. Color code for each Race and Ethnicity Category 14

Figure 7. Visualization of original data on zoom level 5 14

Figure 8. Test-run of Algorithm 1 on zoom level 4 15

Figure 9. Test-run of Algorithm 2 on zoom level 4 17

Figure 10. Test-run of Algorithm 3 on zoom level 4 19

Figure 11. Test-run of Algorithm 4 on zoom level 4 21

List of Tables

Table 1. Remaining Overlapping Pie Charts from Algorithm 1 16
Table 2. Remaining Inclusive Pie Charts from Algorithm 1 16
Table 3. Remaining Overlapping Pie Charts from Algorithm 2 18
Table 4. Remaining Inclusive Pie Charts from Algorithm 2 18
Table 5. Remaining Overlapping Pie Charts from Algorithm 3 20
Table 6. Remaining Overlapping Pie Charts from Algorithm 4 21
Table 7. Summary of Viewpoints 22

Chapter 1 Introduction

Visualization technique is important for analysis and exploration of data [1]. The advantage of visualization technique is that by visualizing information in appropriate manner, it can help to deal with abundance of information as well as to gain more understanding beyond that information. Especially nowadays, there is a lot of information with complex data set, which includes time series, geographic locations, and multiple variable data. We will further address such data set as multi-dimensional data. Multi-dimensional data if being visualized suitably, can provide advance comprehension about the data that may support in deeper analysis.

1.1. Background

Visualization of multi-dimensional data is interesting yet challenging at the same time. The purpose is not only to display the data, but to support user understanding along with the maintenance of proportional ratio of the data itself.

There is a kind of multi-dimensional data that has not only multiple variable data, but also geographical locations and size attribute. Plotting such kind of data into glyphs on 2-D maps might be one of the common analysis methods to be able to represent the spatial distribution and also the dimension of the data itself. In this case, visual cluttering might appear due to unbalance size of the data and size of the location they represent. There could be some overlapping data that cause only several data are visible clearly and some other data are hidden.

1.2. Research Purpose

By taking those mentioned above into consideration, the purpose of our research is to find a technique to reduce overlapping symbols on the visualization of multi-dimensional data while still maintaining the proportional ratio of their original attributes.

1.3. Structure of the Thesis

The structure of this thesis is as follows. In Chapter 2, we review some related works and literatures that became the basis of our research. The challenge of visualizing multi-dimensional data and the usage of Proportional Symbol Map are presented. Some methods to handle visual clutter are also mentioned. Chapter 3 describes some requirements of our technique to achieve our goal and we also elaborate our proposed method and implement some algorithms to handle overlapping symbols based on our findings on Chapter 2 and our requirements. In Chapter 4, we explain our data, experiments, and analysis of the experiments' result. And later, we evaluate our proposed algorithms with some viewpoints that clearly show the comparison of those algorithms. And lastly, in Chapter 5, concluding remarks of our research are given.

Chapter 2 Related Research

2.1. Visualization of Multi-dimensional Data

During past decades, along with the growth of technology and science, the information floods cannot be avoided. That information is not only abundant in number, but also in varieties. It causes special treatment to visualize each variety becomes important to deliver the right insight of information to the reader.

One of the classifications of information is based on the dimension of the data they are represented by. There is a kind of data set that only has one dimension, commonly represented by a Cartesian plane with coordinate points mapped to the x-axes and data values mapped to the y-axes [2]. There is also a kind of data set that has two dimensions, commonly represented by a Cartesian plane with coordinate points mapped to the x- and y-axes, and the data values are mapped to either a set of colors or to the z-axes [2]. Besides that, there is also a kind of data set that has more than two dimensions. It can be addressed as multi-dimensional data (e.g. a data set that has geographical location, time series, size, and multivariate attributes of value).

Methods to visualize multi-dimensional data varies from the usage of information graphics such as charts, histograms, scatter plots [3] through a novel kind of coordinate axes such as Star Coordinate [4]. Kandongan explained the usage of Star Coordinate to visualize multi-dimensional clusters, trends, and outliers. Using Star Coordinates, each multi-dimensional data element is represented by a point, the coordinate axes are arranged on a 2-D surface, and each ax shares the same origin point. Through the author’s findings, Star Coordinate is especially suitable for hierarchical clusters.

Another approach to visualize multi-dimensional data using coordinate axes is parallel coordinate plots (PCP) [5]. PCP lays out coordinates in parallel line, where each data element is denoted as a line going through the coordinate axes at the value it represents on that coordinate. PCP is very suitable for modeling relations, however it requires user understanding about mathematical procedures, therefore it might be not intuitively easy to understand by common readers.

2.2. Usage of Proportional Symbol Map

Proportional symbol map visualizes data on a map by placing certain kind of symbols, which its symbols change size according to the value of the attribute they represent [6]. The size of each symbol on the map is also proportional to each other. In specialization to a data set that has not only multivariate values but also spatial and size attributes, proportional symbol map [7] might be one of some suitable techniques and also naturally easy to understand.

Brewer outlined some usages of proportional symbol map to visualize cancer data [8]. In [8], he shows ratio of prostate cancer deaths for Black and White males on a map in form of pie charts with two different colors which one color represents Black males and another color represents White males. Still in the same paper [8], he shows prostate cancer mortality for only White males using choropleth mapping, with three different kinds of color schemes: sequential with only single hue scheme; sequential with transitional between three hue schemes, and spectral which uses all different colors for each group of data. He also shows prostate cancer mortality for only White males on a map with circle as the symbol which is scaled depends on individual county value [8].

From here we proposed to combine the usage of pie charts and circles with different size to represent multi-dimensional data from a certain area on a map. Pie charts are suitable to visualize multi-dimensional data in one unified symbol, and different sizes of circle are suitable to represent data sizes based on their location on the map. Using this combination, visualization of complex data set that has multivariate values, size, and geographical locations can be achieved with one single view of a map.

2.3. Handling Overlapping Symbol

The usage of symbols that change size according to the location they represent often result in some visual clutters due to display limitation. Dix et al. used random sampling method to simplify the data and demonstrated 2-D zooming technique to show actual density of sampling [9]. Still related to simplifying data, de Berg et al proposed a concept of distribution approximation of a set of points to simplify Dot Maps [10].

Some systems use semantic zoom technique [11] where the representation of data changes according to users' zooming action. This technique can be tricky since the visualization of data varies based on zoom and it requires user interaction. Woodruff et al [12] proposed the idea of having the constant amount of information regardless users' pan and zoom activity [13, 14]. They consider the density of area where there is visual clutter in deciding the suitable representation of data in that area. In [12], they show how their idea tackles visual cluttering by using different graphical representation. In high density area they use smaller polygons, and in low density area they use larger polygons. However, different sizing in the same zoom level might confuse the user in the actual number the symbols represent.

Visual clutter not only happens to a static objects but also dynamic objects. Scheepens et al proposed a method to handle overlapping symbols of moving objects (e.g. data from maritime domain) by dividing the set of all objects into subsets that indicate distribution of several attributes of its objects [15]. Those subsets are scaled based on the density of the objects they represent and will also move accordingly. This method is close to our approach to visualize objects' attributes in one unified symbol like pie charts. However, since we are not dealing with moving objects, we proposed sampling and clustering-based technique are more suitable for our technique.

2.4. Clustering Based on Similarity

The arrangement of dimensions plays an important role in clustering technique. Different order of dimension might cause different result of visualization. Ankerst et al proposed an idea to utilize similarity measurement in rearranging dimensions [16]. Specifically, Ankerst et al use Euclidean distance measure to evaluate global and partial similarity of two dimensions.

In fact, similarity can be measured in various ways. Sawhney et al proposed a method to retrieve images based on their content in terms of color represented as a quadratic form [17]. On the other hand, similarity of objects can also be measured in form of their shapes using normalized Fourier descriptor method [18].

Another form to evaluate similarity between objects is by calculating the cosine measure of their attributes. Wilkinson implemented this measurement to obtain similarity of document retrieval [19]. Egghe et al, from their experiment using cosine measure, found a threshold value that can optimize the visualization of vector space [20]. From here we can grasp that cosine measure is suitable for evaluating similarity in terms of vector space.

Chapter 3 Requirement and Proposed Method

3.1. Requirement

The basic idea of our purpose is to be able to reduce visual clutter and maintaining proportional ratio of data that are being visualized. By taking into account the consideration of multi-dimensional data that has geographical location and size attributes, the straightforward visualization of Proportional Symbol Map, and the possibility of data clustering based on similarity, our approach is to visualize the data with suitable symbols on a map, and merge overlapping symbols that might appear by measuring similarity of the symbols. A certain condition (e.g. suitable threshold) is to be found, and the new values of merged symbols are to be decided.

In order to achieve our purpose, there are some requirements we must consider. The visualization technique must able to show multi-dimensional data at one single view thematic map to ensure the effectiveness of the visualization itself. It must able to reduce overlapping symbols that might appear while still showing proportional ratio of the data throughout the map. And finally, it should be able to maintain user understanding of the data.

3.2. Proposed Method

Our method consists of four steps to be able to decide which overlapping symbols to be merged but still maintain the validity and proportional ratio of the data. Firstly, “Similarity Measurement” is conducted to find better choice of merging data based on similarity. Secondly, “Overlapping Rate” is calculated to define the level of overlapped area between two symbols. The next step is to define the “New Value” of the to-be-merged data. And the last step is the “Algorithm Implementation” to merge the symbols according to the previous steps.

3.2.1. Similarity Measurement

As we mentioned before, in this research we tried to visualize multi-dimensional data that has not only multivariate value attribute, but also geographical location and size attributes. We treat the geographical location attribute as spatial distribution where we will plot the symbols on the map, and the size attribute as the control of proportional ratio of data throughout the map. The multivariate value attribute itself will be treated as an n-dimensional vector which we will evaluate the similarity between one data to another using the calculation of cosine similarity.

Given two n-dimensional vectors ***A*** and ***B***, A_i and B_i are the components of vector ***A*** and ***B*** respectively, the angle between them is θ and the cosine similarity between them, $\cos(\theta)$, is represented using dot product and magnitude as

$$\text{similarity} = \cos(\theta) = \frac{\boldsymbol{A}.\boldsymbol{B}}{\|\boldsymbol{A}\|\|\boldsymbol{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

3.2.2. Overlapping Rate

We chose pie chart as the symbol to represent the multivariate attributes of our data due to its capability to visualize many attributes in one unified symbol. Since the shape of pie chart is circle, we will calculate overlapping rate between two pie charts as intersection area of two circles.

Given two circles with radius r_1 and r_2 and centered at (0,0) and (d,0) are overlapped with each other as follows

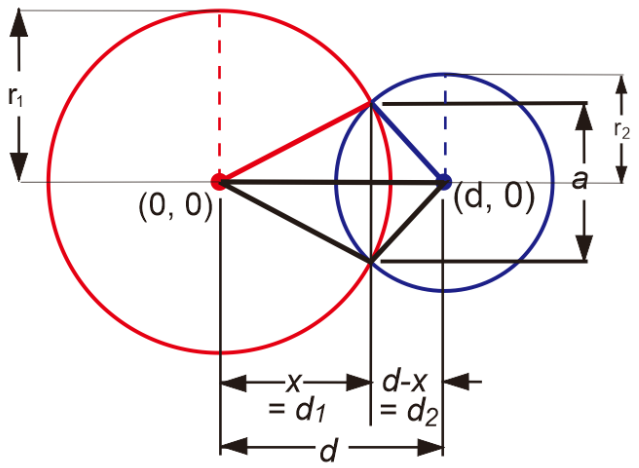

Figure 1. Illustration of Two Overlapped Circles

The equations of the two circles are

$$x^2 + y^2 = r_1^2 \tag{2}$$

$$(x - d)^2 + y^2 = r_2^2 \tag{3}$$

Combining (2) and (3) gives

$$(x-d)^2 + ({r_1}^2 - x^2) = r_2^2 \tag{4}$$

Multiplying through and rearranging gives

$$x^2 - 2dx + d^2 - x^2 = r_2^2 - r_1^2 \tag{5}$$

Solving x results in

$$\mathrm{x} = \frac{d^2 - r_2^2 + r_1^2}{2d} \tag{6}$$

And also solving y from (2) and (6) results in

$$\begin{aligned} \mathrm{y}^2 &= r_1^2 - x^2 = r_1^2 - \left(\frac{d^2 - r_2^2 + r_1^2}{2d}\right)^2 \\ &= \frac{4d^2 r_1^2 - (d^2 - r_2^2 + r_1^2)^2}{4d^2} \end{aligned} \tag{7}$$

Since a = 2y, we can obtain

$$\begin{aligned} \mathrm{a} &= \frac{1}{d}\sqrt{4d^2 r_1^2 - (d^2 - r_2^2 + r_1^2)^2} \\ &= \frac{4d^2 r_1^2 - (d^2 - r_2^2 + r_1^2)^2}{4d^2} \end{aligned} \tag{8}$$

To find the intersection area, which looks like an asymmetric *lens,* we use the formula (9) for the area of circular segment with radius *R'* and height *d'* twice, one for each half part of the *lens* (left part and right part) by taking the heights of them as in expression (10) and (11).

$$\mathrm{A}(R', d') = R'^2 \cos^{-1}\left(\frac{d'}{R'}\right) - d'\sqrt{R'^2 - d'^2} \tag{9}$$

$$\mathrm{d}_1 = x = \frac{d^2 - r_2^2 + r_1^2}{2d} \tag{10}$$

$$\mathrm{d}_2 = \mathrm{d} - \mathrm{x} = \frac{d^2 + r_2^2 - r_1^2}{2d} \tag{11}$$

By using expression (10) and (11) into formula (9), the total area of the *lens,* is

$$\begin{aligned} \mathrm{A} &= \mathrm{A}(\mathrm{r}_1, \mathrm{d}_1) + \mathrm{A}(\mathrm{r}_2, \mathrm{d}_2) \\ &= r_2^2 \cos^{-1}\left(\frac{d^2 + r_2^2 - r_1^2}{2\,d\,r_2}\right) + r_1^2 \cos^{-1}\left(\frac{d^2 + r_1^2 - r_2^2}{2\,d\,r_1}\right) - \\ &\frac{1}{2}\sqrt{(-d + r_2 + r_1)(d + r_2 - r_1)(d - r_2 + r_1)(d + r_2 + r_1)} \end{aligned} \tag{12}$$

Area of circle with radius r_1 and radius r_2 are A_1 and A_2 respectively, defined as

$$\mathrm{A}_1 = \pi {r_1}^2 \tag{13}$$

$$\mathrm{A}_2 = \pi {r_2}^2 \tag{14}$$

And finally, the overlapping rate, *overlap¸* is

$$\mathrm{overlap} = \frac{A}{A_1 + A_2 - A} \tag{15}$$

3.2.3. New Value of Merged Symbols

We proposed the usage of pie charts to represent multi-dimensional data. Therefore, defining the new value of the merged symbols is the same as defining the new value of merged pie charts.

Given two overlapped pie charts which have the same component attributes as follows:

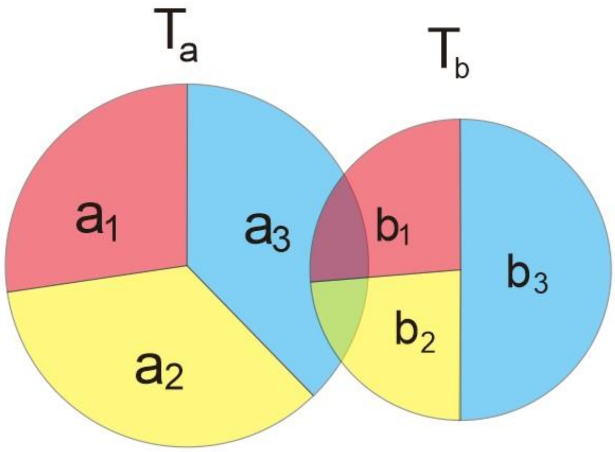

Figure 2. Illustration to Find Similarity of Overlapped Symbols

The left pie chart has total value of T_a and consists of three different attributes which each attribute differs in color, and the value of all attributes are represented as a_1, a_2, and a_3. The right pie chart has total value of T_b, and the value of all attributes are represented as b_1, b_2, and b_3. The total value T_a with three different attributes is expressed as $T_a = a_1 + a_2 + a_3$. The same goes for T_b with three different attributes is expressed as $T_b = b_1 + b_2 + b_3$. Therefore, the total values represented by two overlapped pie charts for n-dimensional data are

$$T_a = a_1 + a_2 + a_3 + \cdots + a_n \tag{16}$$

$$T_b = b_1 + b_2 + b_3 + \cdots + b_n \tag{17}$$

And then to combine those values for a new merged pie chart, we calculate the new value of merged pie chart T_c as

$$T_c = T_a + T_b \tag{18}$$

$$T_c = c_1 + c_2 + c_3 + \cdots + c_n \tag{19}$$

3.2.4. Algorithm Development

In the development of algorithm, we will define the threshold of *similarity* and *overlapping rate* in which the overlapping pie charts will be merged. In our algorithms, the input is *markers,* which is an array of data sort by geographical location, and the output is an array of combined pie charts and other pie charts that do not satisfy the condition of combining pie charts. We divided the usage of those thresholds into two implementations. We will illustrate them as follows:

1. Implementation 1

Given t_1 is the threshold of *similarity* and t_2 is the threshold of *overlap,* the illustration of the usage of these thresholds is in this way:

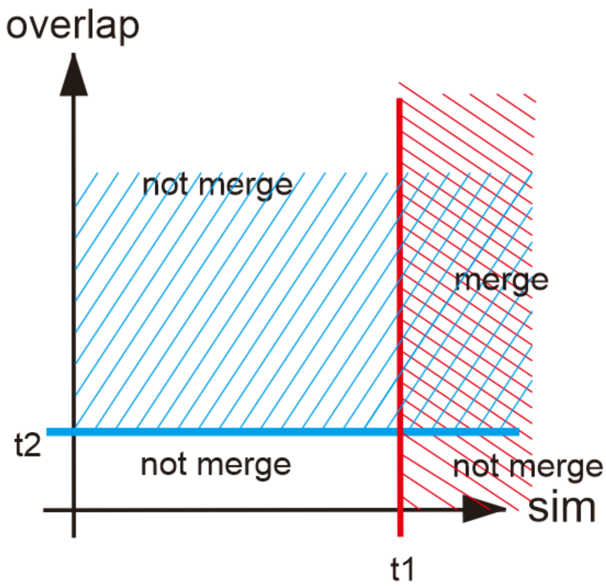

Figure 3. Illustration of Implementation 1

And we also divided the algorithm development into two kinds:

Algorithm 1

In the process of clustering, the original data will not be visible after the overlapping symbols have been combined. This condition is called *lose information* [21]. After being merged, the observer can only see the representation of those symbols. Therefore, in Algorithm 1, we will only merge two pie charts at once to minimize lose information during the merging process.

```
for (k=0; k<markers.length; k++){
    if (markers[k]==null) continue;
    for (l=k+1; l<markers.length; l++){
        if (markers[l]==null) continue;
        //calculate d = distance, r1 = radius of pie charts k, r2 = radius of pie charts l
        if (d<r1+r2){
            //calculate sim = similarity and overlap = overlapping rate
            if ((sim>=t1) && (overlap>=t2)){
                markers[k]=combine(markers[k], markers[l]);
                delete markers[l];
                break; //only combine two pie charts in one time
            }
        }
    }
}
```

Algorithm 2

However, we also would like to observe the result of combining more than two pie charts as long as they fulfill the threshold of *similarity* and *overlap.* The algorithm is

```
for (k=0; k<markers.length; k++){
    if (markers[k]==null) continue;
    for (l=k+1; l<markers.length; l++){
        if (markers[l]==null) continue;
        //calculate d = distance, r1 = radius of pie charts k, r2 = radius of pie charts l
        if (d<r1+r2){
            //calculate sim = similarity and overlap = overlapping rate
            if ((sim>=t1) && (overlap>=t2)){
                markers[k]=combine(markers[k], markers[l]);
                delete markers[l];
            }
        }
    }
}
```

2. Implementation 2

The second implementation still use the threshold t_1 and t_2 but with different type of usage.

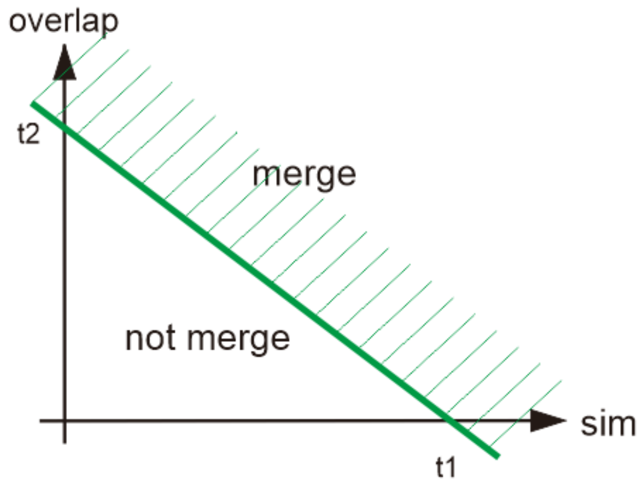

Figure 4. Illustration of Implementation 2

Using this graph, we define the relation of *similarity* and *overlap* as

$$y \geq -\frac{t_2}{t_1}x + t_2 \tag{20}$$

$$overlap \geq -\frac{t_2}{t_1}sim + t_2 \tag{21}$$

$$overlap + \frac{t_2}{t_1}sim \geq t_2 \tag{22}$$

Algorithm 3

Similar to Algorithm 1 above, this algorithm will also merge two pie charts only. The algorithm is in this way

```
for (k=0; k<markers.length; k++){
        if (markers[k]==null) continue;
        for (l=k+1; l<markers.length; l++){
                if (markers[l]==null) continue;
                //calculate d = distance, r1 = radius of pie charts k, r2 = radius of pie charts l
                if (d<r1+r2){
                        //calculate sim = similarity and overlap = overlapping percentage
                        if ((overlap+t2/t1.sim)>=t2){
                                markers[k]=combine(markers[k], markers[l]);
                                delete markers[l];
                                break; //only combine two pie charts in one time
                        }
                }
        }
}
```

Algorithm 4

And lastly, we also tried to merge all pie charts that fulfill the condition in expression (22). The algorithm is as follows

```
for (k=0; k<markers.length; k++){
        if (markers[k]==null) continue;
        for (l=k+1; l<markers.length; l++){
                if (markers[l]==null) continue;
                //calculate d = distance, r1 = radius of pie charts k, r2 = radius of pie charts l
                if (d<r1+r2){
                        //calculate sim = similarity and overlap = overlapping percentage
                        if ((overlap+t2/t1.sim)>=t2){
                                markers[k]=combine(markers[k], markers[l]);
                                delete markers[l];
                        }
                }
        }
  }
```

Chapter 4 Use Case Data and Experiment

4.1. Use Case Data

One data set was employed in this research, which is Race and Ethnicity Data in the United States per 2010. This data was provided by United States Census Bureau website [22]. The Race and Ethnicity category included in the data set are Non-Hispanic White, Hispanic or Latino, Black, American Indian or Alaska Native, Asian, Native Hawaiian or Pacific Islander, and Mixed Race. The data is broken-down by fifty main states in the United States minus Alaska and Hawaii due to their locations which are quite far from the main area of North America thus does not allow overlapping symbols.

Beside the Race and Ethnicity data, we also obtained the population data from QuickFacts page on [22] and Average Latitude and Longitude for US States table from [23]. We use the population data as the reference for the size of the pie charts and the latitude and longitude data as the geographical location on the map where we will put the pie charts onto.

The complete data set of our experiment can be seen in Appendix A.

4.2. Pre-Experiment Activity

First, we tried to visualize the original data onto the map of United States using Google Map API and Javascript programming language.

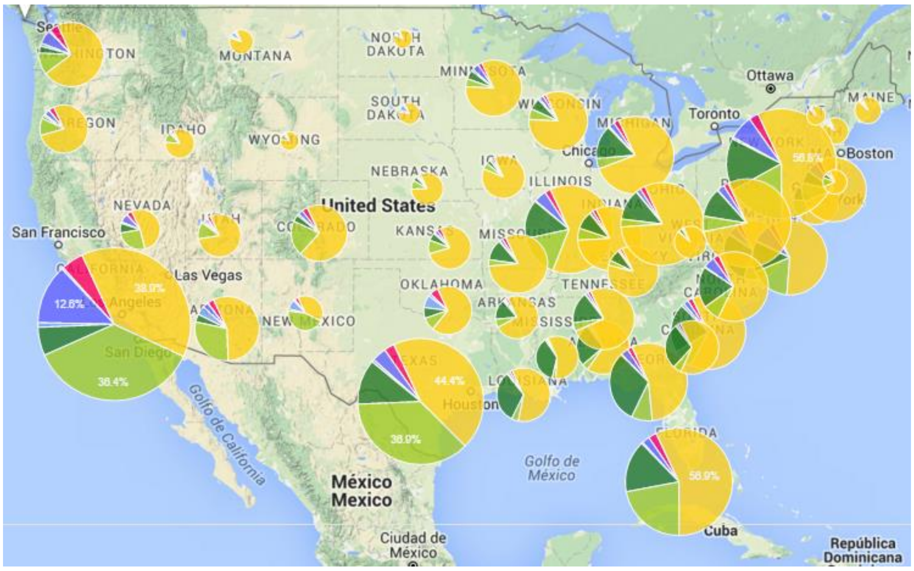

Figure 5. Visualization of original data of Race and Ethnicity in the United States on zoom level 4

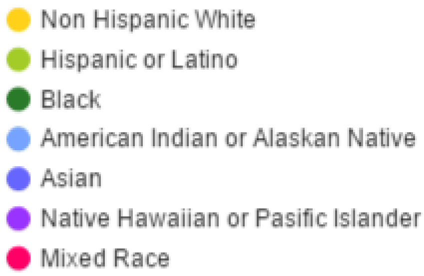

Figure 6. Color code for each Race and Ethnicity Category

From Figure 5 we can see that visualized data is matched to the data in the table in Appendix A. California as the most populous state has the largest pie chart area, followed by Texas at the southern part and New York at the north-eastern part. We can also see that southern part of the United States is mostly dominated by Hispanic or Latino race and Black race (light green and dark green color respectively), while the middle to northern part are dominated by Non-Hispanic White race (yellow color). Using this visualization, we can easily grasp which state is more populous, which race is dominant in which state, and also the ratio of the race itself in each state.

Figure 5 shows the visualization while the zoom level is 4. We can see at this zoom level, there are many overlapping symbols on the right side of the map. It is understandable since the eastern part of the United States is more populous than the western part while there are also more countries with smaller areas (can be confirmed through Appendix A).

In fact, when the zoom level is one level higher, there is much less overlapping symbol on the map, but the display cannot show the whole United States map in one single view.

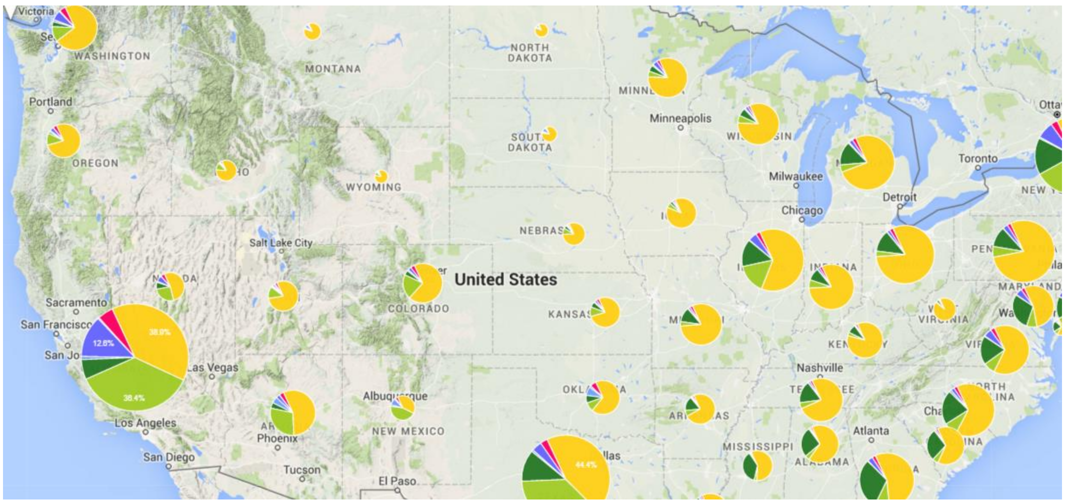

Figure 7. Visualization of original data on zoom level 5

From Figure 7 above we can observe that the southern and eastern part of the United States are outside the display range of the map thus they are invisible. To be able to observe those areas, the users must drag the map to the demanded area.

By considering the fact that on zoom level 5 there are less overlapping symbols but the display cannot show complete data in one single view map, but on zoom level 4, the display is able to show entire data but with many overlapping symbols, we suppose it is important to handle this problem so that our data set can be visualized in one single view map and without many visual clutters.

4.3. Experiment

We test-run our data set to the four algorithms we explained above. In this experiment, we tried to test some cases of *similarity* and *overlap* value, and we found out that most of the *similarity* value of our data was over 0.9 and most of the *overlap* value was over 0.05. Therefore, we decided to use those two values as the threshold of *similarity* (t1) and *overlap* (t2).

Besides that, we also decided some viewpoints to evaluate the effectiveness of all algorithms as well as to compare which algorithm provides the best visualization according to our requirement on section 3.1. Those viewpoints are the total of merged pie charts, the total of overlapped pie charts (means pie charts that are still overlapping even after the algorithm implementation), the total of overlapped area, and the total inclusive pie charts (means pie charts that are not just overlapped, but completely mapped over or under another pie chart). Our program will also calculate these viewpoints.

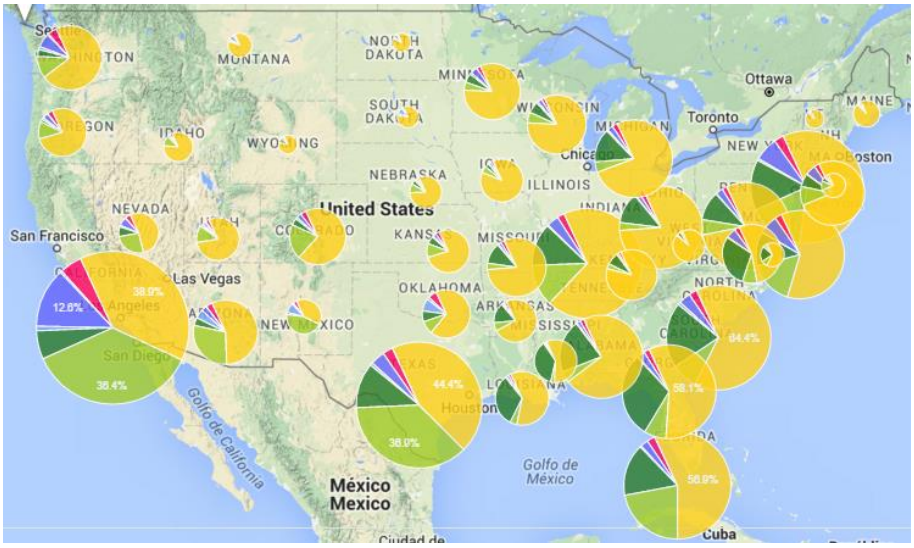

Figure 8. Test-run of Algorithm 1 on zoom level 4

Table 1. Remaining Overlapping Pie Charts from Algorithm 1

No	Overlapping Pie Charts	Similarity	Overlapping Rate	Overlapping Area
1	Florida Georgia-South Carolina	0.952	0.28	1.32
2	Georgia-South Carolina Alabama-Tennessee	0.980	3.73	13.65
3	Georgia-South Carolina North Carolina-Virginia	0.987	6.43	29
4	Alabama-Tennessee Mississippi	0.959	7	13.96
5	Missouri Indiana-Illinois	0.990	3.42	12.74
6	Kentucky Indiana-Illinois	0.986	12.56	39.63
7	Kentucky Ohio	0.998	15.86	63.18
8	Indiana-Illinois Ohio	0.989	2.26	10.18
9	West Virginia Ohio	0.993	9.13	15.43
10	New Jersey-Connecticut New Hampshire-New York	0.999	11.35	52.44
11	New Hampshire-New York Pennsylvania	0.974	8.49	40.3

Table 2. Remaining Inclusive Pie Charts from Algorithm 1

No	Inclusive Pie Charts	Similarity
1	Delaware New Jersey-Connecticut	0.983
2	Rhode Island Massachusetts	0.998
3	Rhode Island New Hampshire-New York	0.979

Algorithm 1 only merged two pie charts at once to minimize the *information lost*. We can see from figure 8 that there are still many overlapping symbols on the north-eastern part of the map. Table 1 shows more detail data of the remaining overlapping and inclusive pie charts.

From table 1 and 2, we can observe that there are many symbols that are actually satisfy our threshold of similarity and overlap, however, since the algorithm only merge two data at once, a lot of symbols remained overlapping on the map.

Algorithm 2 merged all pie charts that fulfill the threshold we have decided. Merged pie charts are Georgia-South Carolina-North Carolina, Alabama-Tennessee, Virginia-Maryland-New Jersey-Connecticut-Massachusetts, Indiana-Illinois-Ohio, New Hampshire-New York, with total fifteen pie charts.

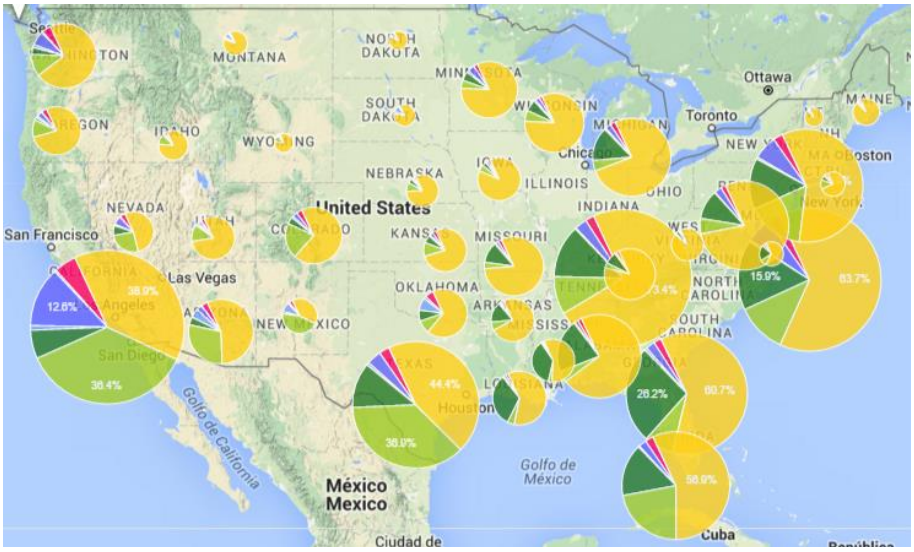

Figure 9. Test-run of Algorithm 2 on zoom level 4

From figure 9, it is clear that although a lot of symbols have been merged, there are still some overlapping symbols left, and even there are some inclusive symbols appear. We supposed that it happens because the *similarity* and/or *overlap* value of those pie charts didn't satisfy our threshold. More detail data about the remaining overlapping and inclusive pie charts can be seen from table 3 and 4.

Table 3 and 4 show that there are some symbols that actually fulfill our thresholds but by some means are not merged. We supposed there are some bugs with our algorithm that made it failed to execute the command or to calculate the *similarity* and/ or *overlap* value of the data.

Table 3. Remaining Overlapping Pie Charts from Algorithm 2

<table>
<tr><th>No</th><th>Overlapping Pie Charts</th><th>Similarity</th><th>Overlapping Rate (%)</th><th>Overlapping Area (pixel)</th></tr>
<tr><td rowspan="2">1</td><td>Florida</td><td rowspan="2">0.963</td><td rowspan="2">4.26</td><td rowspan="2">25.5</td></tr>
<tr><td>Georgia-South Carolina-North Carolina</td></tr>
<tr><td rowspan="2">2</td><td>Georgia-South Carolina-North Carolina</td><td rowspan="2">0.990</td><td rowspan="2">0.7</td><td rowspan="2">3.7</td></tr>
<tr><td>Alabama-Tennessee</td></tr>
<tr><td rowspan="2">3</td><td>Alabama-Tennessee</td><td rowspan="2">0.991</td><td rowspan="2">6.9</td><td rowspan="2">40.7</td></tr>
<tr><td>Indiana-Illinois-Ohio</td></tr>
<tr><td rowspan="2">4</td><td>Alabama-Tennessee</td><td rowspan="2">0.959</td><td rowspan="2">7</td><td rowspan="2">13.96</td></tr>
<tr><td>Mississippi</td></tr>
<tr><td rowspan="2">5</td><td>Indiana-Illinois-Ohio</td><td rowspan="2">0.985</td><td rowspan="2">0.92</td><td rowspan="2">4.37</td></tr>
<tr><td>West Virginia</td></tr>
<tr><td rowspan="2">6</td><td>Virginia-Maryland-New Jersey-Connecticut-Massachusetts</td><td rowspan="2">0.987</td><td rowspan="2">7.53</td><td rowspan="2">48.2</td></tr>
<tr><td>Pennsylvania</td></tr>
<tr><td rowspan="2">7</td><td>Virginia-Maryland-New Jersey-Connecticut-Massachusetts</td><td rowspan="2">0.995</td><td rowspan="2">7.87</td><td rowspan="2">60.18</td></tr>
<tr><td>New Hampshire-New York</td></tr>
<tr><td rowspan="2">8</td><td>Pennsylvania</td><td rowspan="2">0.973</td><td rowspan="2">8.49</td><td rowspan="2">40.3</td></tr>
<tr><td>New Hampshire-New York</td></tr>
</table>

Table 4. Remaining Inclusive Pie Charts from Algorithm 2

<table>
<tr><th>No</th><th>Inclusive Pie Charts</th><th>Similarity</th></tr>
<tr><td rowspan="2">1</td><td>Kentucky</td><td rowspan="2">0.993</td></tr>
<tr><td>Indiana-Illinois-Ohio</td></tr>
<tr><td rowspan="2">2</td><td>Delaware</td><td rowspan="2">0.982</td></tr>
<tr><td>Pennsylvania</td></tr>
<tr><td rowspan="2">3</td><td>Rhode Island</td><td rowspan="2">0.979</td></tr>
<tr><td>New Hampshire-New York</td></tr>
</table>

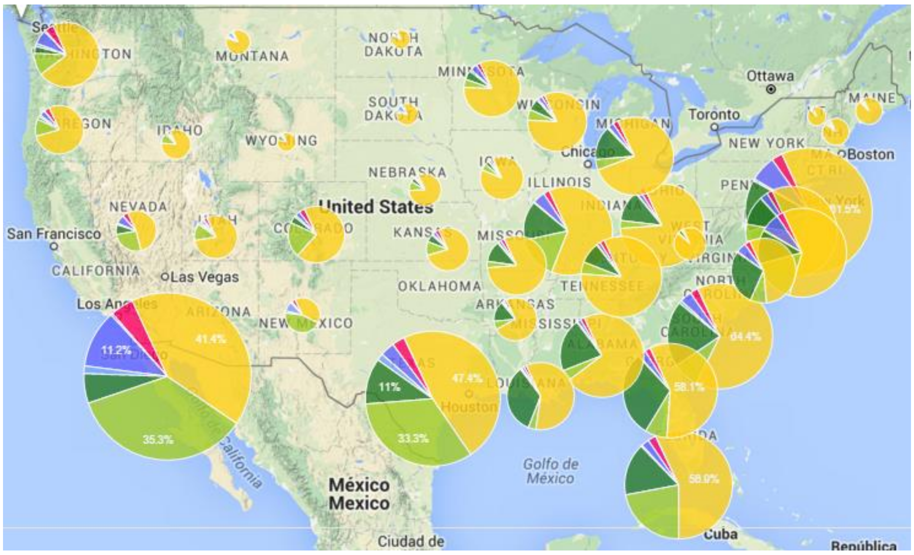

Figure 10. Test-run of Algorithm 3 on zoom level 4

Similar to Algorithm 1, Algorithm 3 only merged two pie charts at one time. Therefore, it is obvious that there are still some overlapping pie charts left on Figure 10. Merged pie charts are Louisiana-Mississippi, Georgia-South Carolina, Alabama-Tennessee, Texas-Oklahoma, Arizona-California, North Carolina-Virginia, Delaware-Maryland, Kentucky-Indiana, New Jersey-Connecticut, Rhode Island-Pennsylvania, and Massachusetts-New York with total 22 combined pie charts. Remaining overlapping and inclusive pie charts can be seen from table 6 and 7 below.

In the test-run of Algorithm 3, there are some pie charts that are not just overlapped but also piled up on each other; they are Delaware-Maryland, New Jersey-Connecticut, Rhode Island, and Massachusetts-New York. On the other hand, there are no inclusive pie charts. However, there are some pie charts that are not overlapping but being merged in the end. Those pie charts are Louisiana-Mississippi (*similarity* = 0.996), Texas-Oklahoma (*similarity* = 0.839) and Arizona-California (*similarity* = 0.942). By looking at Texas and Oklahoma's similarity that does not even fulfill our threshold, we supposed this happens because of the condition in Algorithm 3 "if ((overlap+0.05/0.9*sim)>=0.05)" makes the *similarity* 0.839 satisfy that formula. This case shows that the condition in Algorithm 3 may lead to incorrect merging of pie charts.

And lastly, from Figure 11 we can observe the test-run of algorithm 4. Since the range of threshold is wider (because of the condition "if ((overlap+0.05/0.9*sim)>=0.05)") and more pie charts to be merged due to the merging of all pie charts that fulfill our threshold in one time, we can observe from table 4 that there are less pie charts on the map compared to other algorithms' test-run.

Table 5. Remaining Overlapping Pie Charts from Algorithm 3

No	Overlapping Pie Charts	Similarity	Overlapping Rate (%)	Overlapping Area (pixel)
1	Florida Georgia-South Carolina	0.952	0.28	1.32
2	Georgia-South Carolina Alabama-Tennessee	0.980	3.73	13.65
3	Georgia-South Carolina North Carolina-Virginia	0.987	6.43	29
4	North Carolina-Virginia Delaware-Maryland	0.987	0.91	3.36
5	Delaware-Maryland Massachusetts-New York	0.963	7.55	42.49
6	Delaware-Maryland Rhode Island-Pennsylvania	0.942	17.29	50.93
7	Delaware-Maryland New Jersey-Connecticut	0.958	19.25	43.98
8	Alabama-Tennessee Kentucky-Indiana	0.984	0.36	1.21
9	Kentucky-Indiana Illinois	0.975	3.73	13.65
10	Kentucky-Indiana Ohio	0.999	6.62	22.04
11	Missouri Illinois	0.977	5.64	16.15
12	West Virginia Ohio	0.993	7.73	13.24
13	Ohio Michigan	0.999	2	6.1
14	New Jersey-Connecticut Rhode Island-Pennsylvania	0.98	52.1	129.47
15	New Jersey-Connecticut Massachusetts-New York	0.999	29.32	155.99
16	Rhode Island-Pennsylvania Massachusetts-New York	0.98	37.87	188.1

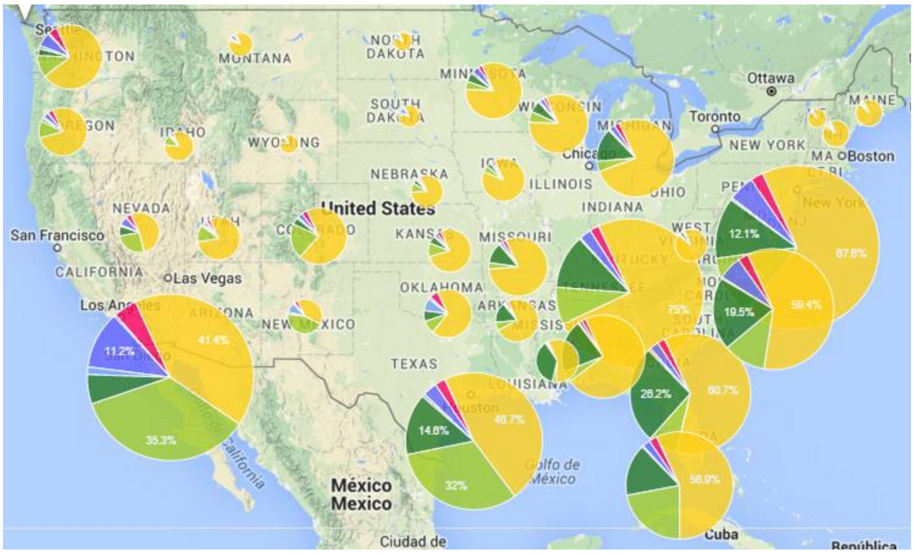

Figure 11. Test-run of Algorithm 4 on zoom level 4

Table 6. Remaining Overlapping Pie Charts from Algorithm 4

No	Overlapping Pie Charts	Similarity	Overlapping Rate (%)	Overlapping Area (pixel)
1	Florida Georgia-South Carolina-North Carolina	0.963	4.26	25.5
2	Georgia-South Carolina-North Carolina Alabama-Tennessee	0.990	0.7	3.7
3	Alabama-Tennessee Mississippi	0.959	7	13.96
4	Alabama-Tennessee Kentucky-Indiana-Illinois-Ohio	0.990	14.08	86.93
5	Georgia-South Carolina-North Carolina Kentucky-Indiana-Illinois-Ohio	0.969	2.55	22.33
6	Kentucky-Indiana-Illinois-Ohio West Virginia	0.988	0.8	4.39
7	Delaware-Virginia-Maryland-New Jersey Connecticut-Rhode Island-Pennsylvania-Massachusetts-New York	0.990	24.93	206.53

From figure 11 we can observe that merged pie charts are Louisiana-Texas, Georgia-South Carolina-North Carolina, Alabama-Tennessee, Arizona-California, Delaware-Virginia-Maryland-New Jersey, Kentucky-Indiana-Illinois-Ohio, Connecticut-Rhode Island-Pennsylvania-Massachusetts-New York with total nineteen pie charts.

Similar to the test-run of Algorithm 3, there are no inclusive pie charts and there are non-overlapping pie charts that are merged, which are Louisiana-Texas (*similarity* = 0.795) and Arizona-California (*similarity* = 0.942). Table 6 shows the detail of remaining overlapping pie charts after the test-run.

4.4. Analysis

As we mentioned before, we will evaluate the effectiveness of all algorithms with some viewpoints, which are the total of merged pie charts, the total of overlapped pie, the total of overlapped area, and the total inclusive pie charts. By observing the results we obtained on 4.3, here is the summary of the viewpoints in Table 7.

If the algorithms were effective to reduce pie charts, we expected they could merge as many as possible pie charts that satisfy our threshold, few remaining overlapped pie charts, small overlapped area, and no inclusive pie charts.

From Table 7, we can see that Algorithm 1 and 3 left more overlapped pie charts compared to Algorithm 2 and 4. It is obvious because Algorithm 1 and 3 only merge two pie charts at once. In case of comparing Algorithm 1 and 3, Algorithm 3 merged more pie charts that Algorithm 1, because of the condition (figure 3) resulted wider range of merging pie chart, and Algorithm 3 also resulted no inclusive pie charts. From here we can conclude that Algorithm 3 satisfy our requirement better than Algorithm 1.

Table 7. Summary of Viewpoints

Algorithm	Total of Merged Pie Charts	Total of Overlapped Pie Charts	Total of Overlapped Area (pixel)	Total of Inclusive Pie Charts
1	12	15	149.66	3
2	15	8	236.91	3
3	22	16	730.68	0
4	19	7	363.34	0

In case of comparing Algorithm 2 and 4, Algorithm 4 merged more pie charts and left less overlapped pie charts and no inclusive pie charts. We can conclude that Algorithm 4 satisfy our requirement better than Algorithm 2.

However, in term of reducing *lost information*, it is clear that Algorithm 2 and 4 might leads to different perception of the pie charts in term of percentage of each element inside the pie charts and the positions after being merged. On the other hand, Algorithm 1 and 3 are not able to simplify the visualization.

In term of comparing Implementation 1 (Algorithm 1 and 2) and Implementation 2 (Algorithm 3 and 4), Implementation 1 resulted inclusive pie charts that actually fulfilled our threshold. On the other hand, although Implementation 2 didn't leave inclusive pie charts, there were some pie charts being merged even though they didn't fulfill our threshold.

In conclusion, the decision of which Algorithm performed the best to visualize our data is closely related to the tendency we want to emphasize. If we want to minimize the *lost information*, Algorithm 3 performed better than Algorithm 1. However, if we want to simplify the visualization by reducing as many as possible pie charts, Algorithm 4 resulted better than Algorithm 2. And finally, the kind of implementation also affect the result. If we want to use Algorithm 3 or 4, we have to examine the bug that resulted in merging of non-overlapping symbol. On the other hand, if we want to implement Algorithm 1 or 2 as they were rooted from the original idea of the threshold, we have to assess what the cause of the existence of inclusive pie charts is.

Chapter 5 Conclusion

5.1. Summary

This research is an attempt to find a novel technique to reduce overlapping symbols on the visualization of multi-dimensional data using Proportional Symbol Map. The experiment is based on a data set of Race and Ethnicity in the United States, provided by the United States Census Bureau. In the methodology, we try to merge overlapping symbols that satisfy our threshold of *similarity* and *overlapping rate.* There are four different algorithms based on the principal of *lost information* and simple visualization implemented to our data and evaluated with some viewpoints.

After our experiment, we find that the effectiveness of the algorithms depend on the user tendency, whether to prioritize the minimum *lost information* or the simple visualization. Some algorithms are better to be used for one principal and some others for another principal. However, our method has been able to reduce overlapping symbols while still maintaining the proportional ratio of the data throughout the map.

5.2. Future Work

This research has still some limitations that are potential to lead for further analysis or more developed methodology. The implementation of the algorithms still showed some bugs that need to be perfected. Some algorithms result inclusive symbols on the visualization although those symbols fulfill our threshold to be merged. Other algorithms merged symbols that are not overlapping. Therefore, we recommend further analysis to improve each imperfection in our algorithms to be able to be used in general cases and with many more data sets.

Our methodology also has not been able to show directly which symbols have been merged. Whereas, this feature is important so that the user can immediately grasp which symbols belong to which states. Besides that, since our methodology is not an animated visualization, we have not been able to show the movement from unmerged symbols to merged ones, while we also think that this feature important so that the user can immediately get the idea how the implementation of merged symbols likes.

Moreover, we only implemented our methodology to one data set. We suppose there might be different result if we implemented more data sets with different characteristic or attributes. This might lead to another conclusion and future works.

Acknowledgement

All the praises be to Allah, without whom I would not be able to go through this journey.

First of all, I would like to express my gratitude to Professor Kazuo Misue who always gives me guidance and valuable advice to my study, and also support during my stay in Japan. Also to all former members of IPLAB and current members of Vislab for all the help in my research and for the friendship they have given to me until now.

My earnest appreciation goes to my husband, Araf Pratamanaim, for his never-ending love and encouragement, and to my beloved parents, Bambang Edyanto, Rita Rusdiani, Oom Somaharja, Imas Permasih, and also to my brother and sisters, Aufan Asidqi, Nisa Permatasari, Intan Oktaviani, for all the prayers, spirit, and care.

In addition, I would like to thank MEXT that provide me the scholarship to pursue my dream of studying in Japan. And also to all fellow MEXT 2013 awardees, especially here in University of Tsukuba; Serign, Leslie, Hanna, Ailyn, Felipe, Marko, Ana, Auchitya, Jaloh, Chitho, Jesse, Shaka, Susana, Vundo, and Wagner, who have become my friends, sisters, and brothers here in Japan.

Last but not least, my special gratitude goes to Indonesian Students Association in Ibaraki, for all the experience, and also to each and every person I met here, that have given special colors and laughter to my life. Thank you for being the part of this journey. Although now we are apart, the memory will last.

Ichinoya Dormitory 18-106, Tsukuba, Japan
December 21st, 2015 – 10:00 AM
Nadine Nandanari

Reference

[1] Keim, D. A.: “Visual Support for Query Specification and Data Mining”, Ph.D. thesis, University of Munich, July 1994, Shaker Publishing Company, 1995.

[2] Johansson, J., Elements of Scientific Visualization Using OpenDX, *Course #AICT-07-08*, 2007.

[3] Harris, R.L., Information Graphics: A Comprehensive Illustrated Reference, p. 448, Oxford, UK: Oxford Press, 1999.

[4] Kandongan, E., Visualizing Multi-Dimensional Clusters, Trends, and Outliers Using Star Coordinates, *Proc. Seventh ACM SIGKDD International Conf. Knowledge Discovery and Data Mining*, pp. 107-116, 2001.

[5] Inselberg, A., “The Plane with Parallel Coordinates,” The Visual Computer, vol. 1, pp. 69-97, 1985.

[6] Definition of proportional symbol. Retrieved on May 2015 from http://support.esri.com/en/knowledgeBase/GISDictionary.

[7] Slocum, T. A., McMaster, R. B., Kessler, F. C., Howard, H. H., “Thematic Cartography and Geovisualization”, Third Edition, pp. 302-326, 009.

[8] Brewer, C., Basic mapping principles for visualizing cancer data using geographic information systems (GIS), *American Journal of Preventive Medicine* 2006, 30:S25-S36.

[9] Dix, A., and Ellis, G., by chance enhancing interaction with large data sets through statistical sampling. In *Proc. of the Working Conf. on Advanced Visual Interfaces*, pages 167–176, 2002.

[10] de Berg, M., Bose, P., Cheong, O., and Morin, P., On simplifying dot Maps, *Comput. Geom. Theory Appl.*, 27(1):43–62, Jan. 2004.

[11] Woodruff, A., Olston, C., Aiken, A., Chu, M., Ercegovac, V., Lin, M., Spalding, M., and Stonebraker, M., DataSplash: A Direct Manipulation Environment for Programming Semantic Zoom Visualizations of Tabular Data, *J. Visual Languages and Computing, special issue on visual languages for end-user and domain-specific programming*, vol. 12, no. 5, pp. 551-571, Oct. 2001.

[12] Woodruff, A., Landay, J., and Stonebraker, M., Constant density visualizations of nonuniform distributions of data, *In Proceedings of the 11th annual ACM symposium on User interface software and technology*, pages 19–28. ACM Press, 1998.

[13] Töpfer, F., and Pillewizer, W., “The Principles of Selection, A Means of Cartographic Generalization,” Cartographic J., 3(1):10-16, 1966.

[14] Frank, A., and Timpf, S., “Multiple Representations for Cartographic Objects in a Multi-Scale Tree - An Intelligent Graphical Zoom,” Computers & Graphics, 18(6):823-829, Nov.-Dec. 1994.

[15] R. Scheepens, H. van de Wetering, and J. J. van Wijk. Non-overlapping aggregated multivariate glyphs for moving objects, *In Proc. IEEE PacificVis*, pages 17–24, 2014.

[16] Ankerst, M., Berchtold. S., and Keim, D. A.,Similarity Clustering of Dimensions for an Enhanced Visualization of Multidimensional Data, *Proc. Int'l Conf. Information Visualization '98*, pp. 52-60, 1998.

[17] Shawney, H., Hafner, J., Efficient Color Histogram Indexing, *Proc. Int. Conf. on Image Processing*, pp. 66-70, 1994.

[18] Wallace, T., Wintz P., An Efficient Three-Dimensional Aircraft Recognition Algorithm Using Normalized Fourier Descriptors, *Computer Graphics and Image Processing*, Vol. 13, pp. 99-126, 1980.

[19] Wilkinson, R., Effective retrieval of structured documents, *In Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Resea7ch and Development in Information Retrieval*, Dublin, Ireland, Association for Computing Machinery, July 1994.

[20] Egghe, L., and Leydesdorff, L., The relation between Pearson's correlation coefficient r and Salton's cosine measure, *Journal of the American Society for Information Science and Technology*, 60(5), 1027–1036, 2009.

[21] Andritsos, P., and Tzerpos, V., Software clustering based on information loss minimization. *In Proc. WCRE*, pages 334–344. IEEE, 2003.

[22] Race and Hispanic Origin. Retrieved on July 2015 from http://www.census.gov/quickfacts/table/PST045215/00.

[23] Average Latitude and Longitude for US States. Retrieved January 2015 from http://dev.maxmind.com/geoip/legacy/codes/state_latlon/.

Appendices

Appendix A

Race and Ethnicity Data of United States per 2010 breakdown by States, sort by population in descending manner

States	Latitude	Longitude	Population	Non-Hispanic White	Hispanic or Latino	Black	American Indian or Alaskan Native	Asian	Native Hawaiian or Pacific Islander	Mixed Race
California	37.17	-119.746	38041430	40.1	37.6	6.2	1	13	0.4	4.9
Texas	33.106	-98.6475	26059203	45.3	37.6	11.8	0.7	3.8	0.1	2.7
New York	44.1497	-74.9384	19570261	58.3	17.6	15.9	0.6	7.3	0	3
Florida	28.8333	-81.717	19317568	57.9	22.5	16	0.4	2.4	0.1	2.5
Illinois	42.3363	-89.0022	12875255	63.6	15.8	14.5	0.3	4.6	0	2.3
Pennsylvania	41.5773	-77.264	12763536	79.5	5.7	10.8	0.2	2.7	0	1.9
Ohio	41.3736	-82.7755	11544225	81.1	3.1	12.2	0.2	1.7	0	2.1
Georgia	32.9866	-83.6487	9919945	55.9	8.8	30.5	0.3	3.2	0.1	2.1
Michigan	44.3504	-84.5603	9883360	76.6	4.4	14.2	0.6	2.4	0	2.3
North Carolina	36.6411	-79.8431	9656401	65.3	8.4	21.5	1.3	2.2	0.1	2.2
New Jersey	40.314	-74.5089	8864590	59.3	17.7	13.7	0.3	8.3	0	2.7
Virginia	38.768	-78.2057	8185867	64.8	7.9	19.4	0.4	5.5	0.1	2.9
Washington	48.3917	-121.571	6897012	72.5	11.2	3.6	1.5	7.2	0.6	4.7
Massachusetts	43.2373	-71.5314	6646144	76.1	9.6	6.6	0.3	5.3	0	2.6
Arizona	36.2543	-111.388	6553255	57.8	29.6	4.1	4.6	2.8	0.2	3.4
Indiana	40.8647	-86.2604	6537334	81.5	6	9.1	0.3	1.6	0	2
Tennessee	36.7449	-86.7489	6456243	75.6	4.6	16.7	0.3	1.4	0.1	1.7
Missouri	39.4623	-92.302	6021988	81.0	3.5	11.6	0.5	1.6	0.1	2.1
Maryland	40.0724	-76.7902	5884563	54.7	8.2	29.4	0.4	5.5	0.1	2.9
Wisconsin	45.2563	-89.6385	5726386	83.3	5.9	6.3	1	2.3	0	1.8

Minnesota	46.7326	-93.9196	5379139	83.1	4.7	5.2	1.1	4	0	2.4
Colorado	40.8497	-105.327	5187582	70.0	20.7	4	1.1	2.8	0.1	3.4
Alabama	34.8974	-86.8073	4822023	67.0	3.9	26.2	0.6	1.1	0	1.5
South Carolina	34.8191	-80.9066	4723723	64.1	5.1	27.9	0.4	1.3	0.1	1.7
Louisiana	32.1801	-91.8749	4601893	60.3	4.2	32	0.7	1.5	0	1.6
Kentucky	38.669	-84.6514	4380415	86.3	3.1	7.8	0.2	1.1	0.1	1.7
Oregon	45.5672	-122.127	3899353	78.5	11.7	1.8	1.4	3.7	0.3	3.8
Oklahoma	36.5376	-96.9247	3814820	68.7	8.9	7.4	8.6	1.7	0.1	5.9
Connecticut	42.5834	-72.7622	3590347	71.2	13.4	10.1	0.3	3.8	0	2.6
Iowa	43.0046	-93.214	3074186	88.7	5	2.9	0.4	1.7	0.1	1.8
Mississippi	33.7673	-89.6812	2984926	58.0	2.7	37	0.5	0.9	0	1.1
Arkansas	35.9513	-92.3809	2949131	74.5	6.4	15.4	0.8	1.2	0.2	2
Kansas	39.5111	-96.8005	2885905	78.2	10.5	5.9	1	2.4	0.1	3
Utah	40.1135	-111.854	2855287	80.4	13	1.1	1.2	2	0.9	2.7
Nevada	40.4199	-117.122	2758931	54.1	26.5	8.1	1.2	7.2	0.2	4.7
New Mexico	35.8375	-106.237	2085538	40.5	46.3	2.1	9.4	1.4	0.1	3.7
Nebraska	42.1289	-98.2883	1855525	82.1	9.2	4.5	1	1.8	0.1	2.2
West Virginia	39.468	-80.9696	1855413	93.2	1.2	3.4	0.2	0.7	0	1.5
Idaho	44.2394	-114.51	1595728	84.0	11.2	0.6	1.4	1.2	0.1	2.5
Maine	45.6074	-69.3977	1329192	94.4	1.3	1.2	0.6	1	0	1.6
New Hampshire	43.4108	-71.5653	1320718	92.3	2.8	1.1	0.2	2.2	0	1.6
Rhode Island	42.6772	-71.5101	1050292	76.4	12.4	5.7	0.6	2.9	0.1	3.3
Montana	46.9048	-110.326	1005141	87.8	2.9	0.4	6.3	0.6	0.1	2.5
Delaware	40.3498	-75.5148	917092	65.3	8.2	21.4	0.5	3.2	0	2.7
South Dakota	45.2853	-99.4632	833354	84.7	2.7	1.3	8.8	0.9	0	2.1
North Dakota	48.5362	-99.793	699628	88.9	2	1.2	5.4	1	0	1.8
Vermont	45.0407	-72.7093	626011	94.3	1.5	1	0.4	1.3	0	1.7
Wyoming	43.7475	-107.209	576412	85.9	8.9	0.8	2.4	0.8	0.1	2.2