Inferring Interests from Mouse Clicks: A Data Modeling Approach To Predict Users' Interests With Visualization Systems

Surina Puri[1], Dr. Alvitta Ottley[2]
[1]Georgia Institute of Technology, Atlanta, GA 30332, USA
[2]Washington University in St. Louis, St. Louis, MO 63130, USA

Abstract - Visualizations help tackle a class of problems that neither the computer nor the human can solve on their own. The computer performs complex computations and displays a visual representation of the data so that the human can reason and make judgments. However, the computer's role is typically limited once the data is displayed. We hypothesized that we can improve human-computer collaboration by making predictions about users' interests. In our research, we performed preprocessing by using K Means clustering to develop models of the data. We developed an algorithm to analyze users' input and identify the data cluster that best represents users' interactions or interest. To evaluate the algorithm, we conducted a user study and recorded users' mouse clicks as they interacted with a crime map. We analyzed users' mouse clicks and demonstrated an overall success of 80% at predicting users' interests. The results of these work lay the foundation for developing interfaces that would adapt to assist the users in their tasks and thus create predictive visualizations.

Index Terms - Visualization, Machine Learning, Clustering, Mouse Clicks

1. INTRODUCTION

Visual analytics is the science of analytical reasoning supported by interactive visual interfaces[1]. There is abundance of data in today's world. Visual analytics provide a comprehensive way to interpret this data visually. Not only are visual analytics wide spread in the field of data mining, but also play an important role in day to day online interactions. For example, while browsing for a house on a map based search website, the users are performing visual analytics as they explore and interact with the visualization.

The human-computer collaboration of visual analytics today is passive because, after displaying the visualization, the role of computer is limited and it is left to the user to interact with the visualization and understand the data. This collaboration in visual analytics can be made proactive by enabling the computer to predict users' interest based on preprocessed data modeling. This proactive approach would create predictive visualizations, assist users in their tasks and provide a faster way of visual analysis.

2. RELATED WORK

For the computer to assist the users with their visual analytics tasks, it is important for the computer to understand the users' reasoning process. *Finding Waldo* research focused on this aspect[2]. The researchers conducted an experiment in which participants interacted with a visualization to find Waldo. They applied machine learning algorithms to analyze users' interaction data. The analysis could explain the users' task performance and predict whether a user will be fast or slow at completing the task. From the analysis of users' interaction data, they were also able to infer users' personality traits, including locus of control, extroversion and neuroticism. Thus, by their study, they were able to establish that users interactions can provide information to the computer about the user. In our research, we aim to make the computer assist the user based on its understand of the users' interactions. We hypothesized that we can improve human-computer collaboration by making predictions about users' interests. There are a lot of aspects of a users' interactions, for example, mouse clicks, eye movement, etc. We primary study the users' mouse clicks as their interactions.

Fu, Eugene Yujun, et al [3], in their research analyze users' mouse clicks and aim to predict the users' next next clicks. The two prediction models used in this research are: a model that considers only historic mouse activity sequences, and a multimodal model that utilizes mouse interaction signals and features extracted from mouse trajectory and clicking events. The results indicated that they can dynamically learn a multimodal model that can effectively predict users' next activity from past interaction sequences and mouse interaction signals. Our research shares the same goal of predicting users' interests/intentions, however, we adopt a data modeling approach for our study. Instead of building models of the users' interaction, we build models of the data. We use K Means clustering algorithm to build these models. We developed an algorithm to analyze users' input and identify the data cluster that best represents users' interactions or interest.

3. METHODS

The methods comprised of two major steps; first, developing the algorithm and second, evaluating it.

3.1 ALGORITHM

The Algorithm comprises of two parts; Preprocessing and Analyzing. Python scikit-learn machine learning package was used to preprocess data and perform clustering.

3.1.1 PREPORCESSING:

Step 1: Feature Selection

Remove features with low variance
Remove features that are not visible to the user

After performing the first step of the algorithm the number of features reduced from 20 to 4. Thus, possible interests/subsets of features reduced from 1,048,575 to 15. The featuresAll list comprises of these 4 features.

featureAll = [“Category of Crime”, “Neighborhood”, “Date”, “Time”]

Step 2: Modeling data - building and storing clusters

```
for featureSubset in all possible subsets of featuresAll:
   perform kMeans clustering on featureSubset
   for i in range(number of clusters):
      store featureSubset, points, centroid
```

The goal of step two is to perform clustering by taking a subset of features at a time and storing all the clusters. We saved the results from step two in the following JSON format, where k is the number of clusters :

```
“featureSubset#0” : {
     “features”:[“Category of Crime”]
          “cluster#0”: {
          “pointsInCluster” :
          “centroidOfCluster”:
          …
          “cluster#k”: {
          “pointsInCluster” :
          “centroidOfCluster”:
…
“featureSubset#14” : {
     “features”:[“Category of Crime”, “Neighborhood”,
                 “Date”, “Time”]
          “cluster#0”: {
          “pointsInCluster” :
          “centroidOfCluster”:
          …
          “cluster#k”: {
          “pointsInCluster” :
          “centroidOfCluster”:
```

3.1.2 ANALYZING

Step 3: Processing Input by Indexing Into Stored Clusters

```
userClicks = points that the user clicked on
for i in userClicks:
   for each featureSubset:
      for each cluster:
         confidence = percentage of points of till i that belong in this cluster
         store featureSubset#, features, cluster, confidence
            store featureSubset# and cluster# with max confidence
winningCluster = featureSubset# and cluster# with max confidence

#breaking ties
for i in userClicks:
   currMax = [cluster/s with max confidence]
   if len(cuurMax) > 1:
      for j in currMax:
         distPointCentroid = calculate mean dist. between all points and respective centroid
      break ties by choosing the cluster with the lowestdistPointCentroid
```

We performed step 3 for all tasks and for all the users that completed that task.

Step 4: Retrieving Users’s interest

```
userInterest = “feature” in the winning featureSubset#
userInterstPoints = “points” in the winning cluster#
```

“featureSubset#” and “cluster#” represent the users’ interest and specific cluster/points that the user is interested in that featureSubset. We retrieved this information by indexing into the stored JSON file.

3.2 EVALUATION

To test whether our algorithm can accurately predict the users’ interests, we conducted a user study and recorded users’ mouse clicks as they interacted with a crime map.

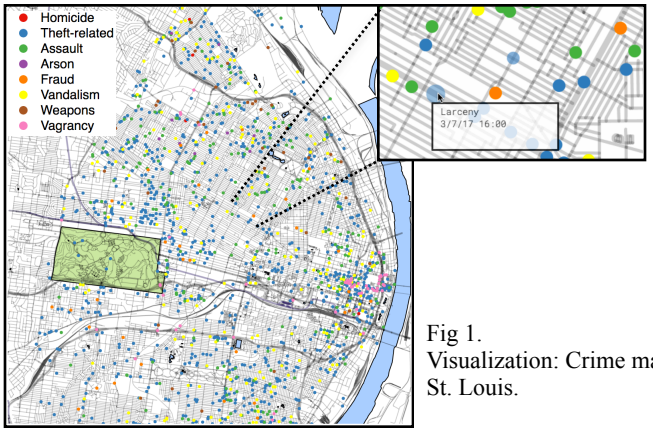


Fig 1.
Visualization: Crime map of St. Louis.

The visualization that the users interacted with in the user study is shown in Fig 1. The data used in this visualization is taken from St. Louis Metropolitan Police

Department[4]. This data set contained 20 features and 1951 instances. It shows crimes that occurred in the city during march 2017. The users can zoom into the map and click on points the see details for category of crime, time and date.

The users were asked to complete 6 tasks. Each task required them to answer a question. To answer each question, the users were required to interact with the map and click on certain points with a common underlying interest. We define users' interest as a subset of all the features of the data set. For example, if the data set comprises of 'crime' and 'date', then there can be 3 subsets/interest: 'crime',' date' and 'crime+date'

Following are the 6 tasks used in the user study:

Table 1

Interest	Task ID	Task
Category of Crime	1	Out of all the cases of Homicide, one case differs from the other cases with regard to time. What is the time of that case?
Category of Crime	2	How many cases of arson occur during PM?
Category of Crime + Neighborhood	3	There are four types Theft-Related crime in the red shaded region: Larceny, Burglary, Robbery and Motor Vehicle Theft. Count the number of cases of Robbery in the red shaded region.
Category of Crime + Neighborhood	4	There are two types of Assault: Aggravated and Non-Aggravated assault. Count the number of Non-Aggravated Assault in the red shaded region
Neighborhood	5	Count the number of crimes that occur during 7:00 AM - 12:30 PM in the red shaded region.
Neighborhood	6	Count the number of crimes during AM in the red shaded region.

For tasks 3-6, different red regions were highlighted. The red regions were randomly drawn over some neighborhoods. The map of St. Louis was divided into 88 neighborhoods. The main neighborhoods that fall into the red shade region for theses task are shown in Table 2.

Table 2

Task ID	Red handed Region
3	Neighborhood IDs: 15, 22-27
4	Neighborhood IDs: 46-52, 68, 78
5	Neighborhood ID: 39
6	Neighborhood IDs: 35-37

Users were asked to complete these tasks in random order. Users' responses for each question and their mouse clicks for were recorded in the database. After completing the 6 tasks, the users also completed a spatial ability survey and personality survey. Their responses for these questions were recoded in the database. At the end, users were asked to provide demographic information and inform if they are color blind.

Visualization of the crime map of St. Louis was developed using D3.js. Amazon Mechanical Turk platform was used to develop the user study and store users' interaction information. This user study was hosted on Amazon Mechanical Turk. 30 users participated in the user study.

We filtered users based on number of points clicked for each question. After filtering, the number of users for the 6 tasks were 17, 17, 20, 20, 19, 20 respectively.

4. RESULTS

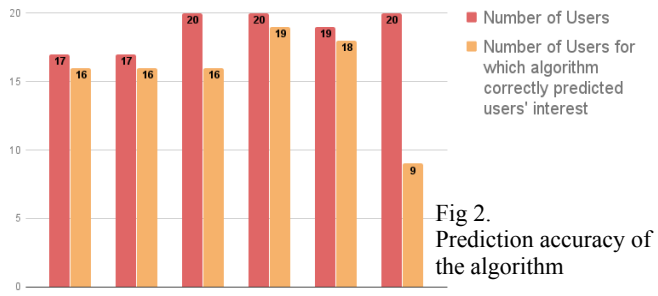
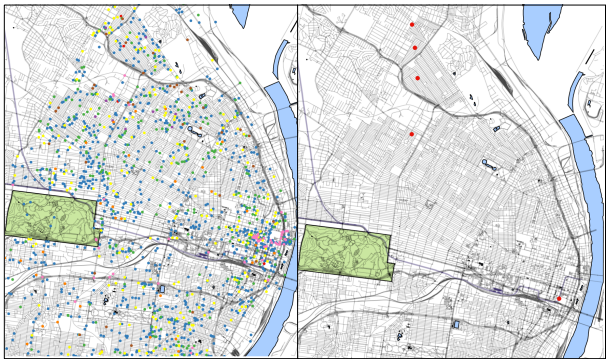


Fig 2.
Prediction accuracy of the algorithm

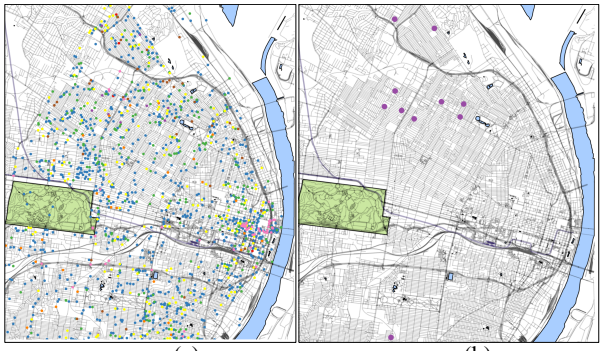
The prediction accuracy of the algorithm is shown in Fig 2. For the first five tasks, it was able to accurately predict users' interests with 80-95 percent accuracy.



(a) (b)

Fig 3. Results from algorithm for Task 1:
(a) Visualization for task 2 - users were expected to click on red points.
(b) Points predicted by the algorithm

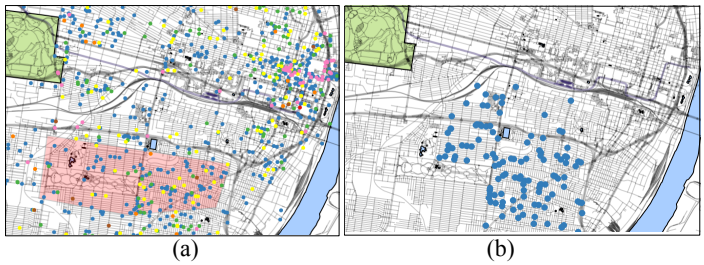
From Fig 3., for task 1, the predicted interest with the highest confidence was Category of Crime. The predicted cluster in this subset contained all points that belong to Homicide. The cluster with the highest confidence correctly represents the users' interest.



(a) (b)

Fig 4. Results from algorithm for Task 2:
(a) Visualization for task 2 - users were expected to click on purple points.
(b) Points predicted by the algorithm

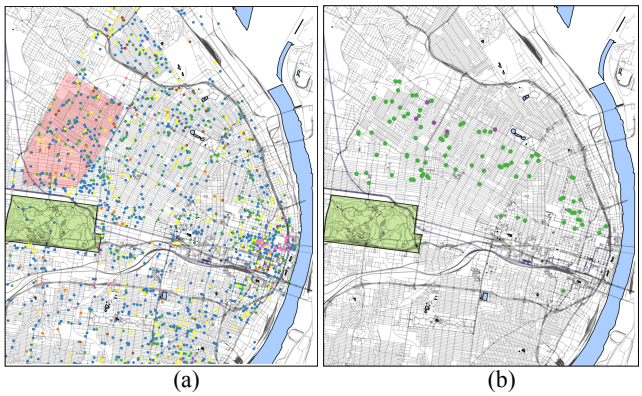
From Fig 4., for task 2, the predicted interest with the highest confidence was Category of Crime. The predicted cluster in this subset contains all points that belong to Arson. The cluster with the highest confidence correctly represents the users' interest.



(a) (b)

Fig 5. Results from algorithm for Task 3:
(a) Visualization for task 3 - users were expected to click on blue points.
(b) Points predicted by the algorithm

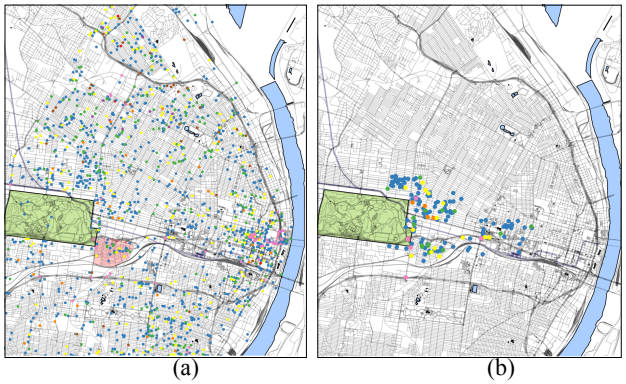
From Fig 5., for task 3, the predicted interest with the highest confidence was Category of Crime and Neighborhood. The predicted cluster in this subset contains all points that belong to Theft in neighborhood IDs 22-31. The cluster with the highest confidence correctly represents the users' interest.



(a) (b)

Fig 6. Results from algorithm for Task 4:
(a) Visualization for task 3 - users were expected to click on green points in the red shaded region.
(b) Points predicted by the algorithm

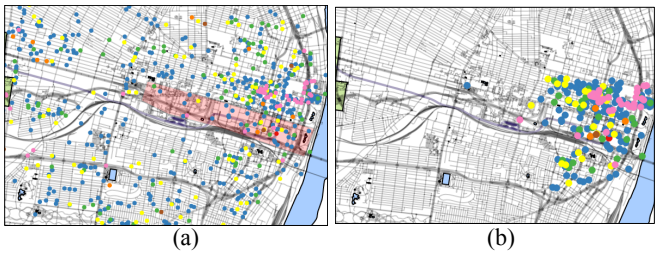
From Fig 6., for task 4, the predicted interest with the highest confidence was Category of Crime. The predicted cluster in this subset contains points that belong to Arson in neighborhood IDs 46, 48, 50-62. The cluster with the highest confidence correctly represents the users' interest.



(a) (b)

Fig 7. Results from algorithm for Task 5:
(a) Visualization for task 3 - users were expected to click on all colors of points in the red shaded region.
(b) Points predicted by the algorithm

From Fig 7., for task 5, the predicted interest with the highest confidence was Neighborhood. The predicted cluster in this subset contains points that belong to all categories of crime in neighborhood IDs 37-40. The cluster with the highest confidence correctly represents the users' interest.



(a) (b)

Fig 8. Results from algorithm for Task 6:
(a) Visualization for task 3 - users were expected to click on all colors of points in the red shaded region.
(b) Points predicted by the algorithm

From Fig 8., for task 6, the predicted interest with the highest confidence was Neighborhood. The predicted cluster in this subset contains points of all categories of crime in neighborhood IDs 33-36. However, the cluster with the highest confidence did not correctly represent the users' interest. Fig 4. (b) shows the cluster with the second highest confidence. This cluster correctly represents users interest. The results of task 6 show a limitation of our algorithm which is discussed in the next section .

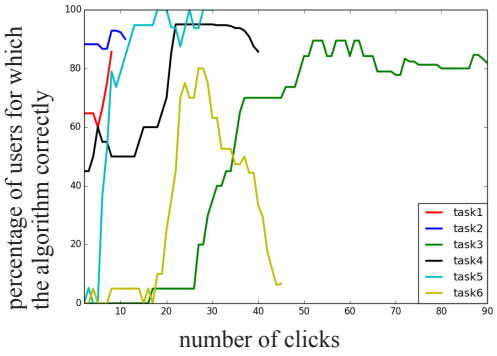


Fig 9. Prediction accuracy of the Algorithm at every click

Fig 9 shows percent of users for which algorithm predicted interests accurately VS average number of clicks for that task

DISCUSSION

Fig 2. shows that the algorithm generally works correctly 80% of the times. However, it does not perform well on task 6. For this task, the cluster that received the highest confidence was a cluster that contained all five homicide points. This cluster is not the correct interest for task 6. The cluster that received the second highest confidence is the correct prediction. Fig 8(b) shows the points in the second highest clustering.

Fig 3, 4,5,6, and 7 show that along with identifying the users' interest/correct subset of features, the algorithm was also correctly able to identify the exact cluster (points in Arson, points in Theft and a specific neighborhood, etc.).

Results of task 6 highlights a limitation of our algorithm. The algorithm is biased towards smaller sized clusters. One future direction to rectify this problem is reevaluating how we calculate confidence. Instead of calculating the confidence based on the cluster that contains the maximum number of points from the user input, the increase/decrease in confidence at each click can also be used to identifying the best cluster for the task.

6. FUTURE WORK

This algorithm can be used for predicting users' interests while they are performing a task. Fig 9. shows how the prediction accuracies generally increase as the number of clicks increase. This can be used to establish the minimum number of clicks required for each task before the algorithm can make accurate real time prediction.

There are a lot of visualizations where this algorithm can prove to be helpful. For example, a map based house search website. These website have an interface similar to the one used in this research. We can also generalize this algorithm to work with various other types of visualizations.

7. CONCLUSTION

The goal for this project was to establish is if we can make human computer collaboration of visual analytics efficient by modeling data and making predictions about users' interests based on their mouse clicks. We developed an algorithm to perform preprocessing and clustering on the data and to analyze the users' mouse clicks. Upon evaluation of the algorithm, we found that not only were we able to identify the users' interest with 80% of the time, we were also able to identify what points/sub features they are interested in. The results shows that, generally, as the number of clicks increase, the accuracy of prediction of interests also increases. This algorithm can be used in real-time to develop interfaces that would adapt to assist the users in their tasks and thus create predictive visualizations.

9. REFERNCES

[1] Thomas, J., Cook, K.: Illuminating the Path: Research and Development Agenda for Visual Analytics. IEEE-Press (2005)

[2] Brown E. Ottley A. Zhao H. Lin Q. Souvenir R. Endert A. Chang R. IEEE Transactions on Visualization and Computer Graphics(2014)

[3] Eugene Yujun Fu, Tiffany C.K. Kwok, Erin You Wu, Hong Va Leong, Grace Ngai and Stephen C.F. Chan. "Your Mouse Reveals Your Next Activity: Towards Predicting User Intention from Mouse Interaction". In 41st Annual IEEE Computer Software and Applications Conference (COMPSAC 2017), Torino, July 4-8 2017.

[4] Data Source: St. Louis Metropolitan Police Department Downloadable Crime Files

10. ACKNOWLEDGEMENTS

The authors would like to thank Dr. Sanmay Das for his valuable feedback.