Tourists Mobility on Social Media

M. Hort, C. Zhang, K. Shingjergji, M. Ignéczi
Dr. Mena Habib

January 26, 2018

Abstract

Social media is becoming a powerful data source for analytics in many domains. This is owing to an increase in quantity and quality of extractable data. Quantitatively, more people seem to find themselves as social media users. Quality wise, users tend to share more personal information and experiences which makes their *profile* more informative. Tourism is one of the main industries to utilize social media, as users posts useful information as they are traveling. The purpose of this research is to identify tourists from social media data and analyze their movement in the Netherlands. This study can be divided into three tasks. Firstly, the detection of tourists, followed by extracting their movement and lastly creating route recommendations for tourists.

1 Introduction

International tourist arrival in Europe has almost doubled in 20 years, increasing from 261.5 million people in 1990 to 476.6 million people in 2010(*BBC-GSCE Bitsize: The growth of Tourism*, n.d.). Tourism is an influence most countries in the world are subject to, benefit from, and need to deal with. While inherently tourism has a positive effect, countries with open borders and less control over the flow of people are having more and more trouble with an increasing number of tourists, leading to overcrowded areas and constantly sold out accommodations. This is not only a problem to countries and cities themselves, but also to tourists who prefer having an insight on how other tourists travel.

Accordingly, the last years have shown a trend of continuous growth in the use of social media. Social media became one of the main communication channels for tourists, to either stay in contact with acquaintances or share their experiences. As a result of this trend, an increasing amount of information about tourists can be obtained from social media and used for research. One field of research related to tourists on social media is the study of tourist mobility. To analyze and research tourist mobility from social media posts, one needs to collect posts from an individual or a group of tourists and extract locations from these posts. This presents a couple of interesting problems, such as identifying tourists solely based on their social media activity or retrieving locations from posts.

In accordance to that, Twitter introduced geotagged tweets in 2009, which enabled the matching of tweets with their location by latitude and longitude (Sarver, 2009). The usage of geotagged tweets increased to a total of approximately 0.77% of 400 million tweets made in 2012. Thereby, 60% of tweets have been posted with mobile devices, feasible for geotagging (Fujita, 2013).

For this research, Twitter will be used to analyze tourist movement within the Netherlands, as Twitter provides easily accessible public data, geotags and profile information. Tweets made in the Netherlands will be retrieved and explored to answer the main research question: ***Can tweets or social media posts map out tourists movement in a country?*** As this question has both a functional and a research focused side, once the mapping is done, analysis needs to be conducted on the result, leading to further research questions:

- Are there popular routes or travel patterns within the country?
- Can routes be suggested to tourists based on other tourists‘ preferences?

Secondarily, this research will explore possible solutions on the question of *whether tourists can be identified based on social media posts*. While this secondary question does not explore tourist mobility, it is an essential part of dealing with tourists on social media, and can prove beneficial to researchers aiming to explore social media and tourism.

2 Related Work

The field of tourism has attracted many researchers either for the purpose of building recommender systems

(Lucas et al., 2013) or to analyze tourists spatial behavior(Edwards & Griffin, 2013). Therefore, data can be retrieved from surveys, GPS, as well as social media platforms, that become more and more important with an increasing usage.

Furthermore, social media has been the focus of various researchers, such as Fujita (Fujita, 2013) who described a collection and visualization of geographical Twitter data. Additional research concerning tourists on social media has been conducted by Munar et al. (Munar & Jacobsen, 2014) who analyzed motivations for sharing tourism experiences through social media. This field of research gets extended by the consideration of geotagged social media posts. Among others, Van Canneyt et al. (Van Canneyt, Schockaert, Van Laere, & Dhoedt, 2011) proposed a system to recommend attractions to tourists based on their location, retrieved from social media posts. Panizio (Panizio, 2015) examines Twitter data in order to get insights on aggregation patterns of tourists near Points of Interest, as Cranshaw et al. (Cranshaw, Schwartz, Hong, & Sadeh, 2012) have shown a correlation between user locations and the proximity to Points of Interest. For this purpose, tweets made in Amsterdam have been analyzed. Lee et al. (Lee, Wakamiya, & Sumiya, 2013) proposed an approach of aggregating patterns on a larger granularity by characterizing cities in Japan based on tweets. Global movement patterns, deduced from geotagged tweets, have been analyzed by Haweka et al. (Hawelka et al., 2014). The application of geotagged tweets reaches as far, that real-time event detections systems have been introduced, by Sakaki et al. (Sakaki, Okazaki, & Matsuo, 2010) to detect earthquakes .

In this paper, we extend upon a heuristic tourism detection method (Habib & Krol, 2017), recognizing tourists based on the time difference between their first and last tweet. This approach is extended by using co-training to detect tourists by among others analyzing the content of their tweets. Different visualization models will be created in order to explore tweets and tourist movement. Lastly, methods to suggests travel routes to tourists will be presented.

3 Concepts and Approach

The distinction of tourists from residents formulates the basis of the following project. Therefore, multiple approaches will be followed and evaluated. Firstly, tourists will be distinguished from residents by comparing the date of their first and last tweet, assuming that residents tweet within a larger timespan. Timespans of various length will be examined. Furthermore, the mobility and movement of tourists will be analyzed. This analysis will be carried out focusing on tweets made in 2016.

Another approach to detect tourists is to consider the content of their tweets. Therefore a model can be deployed, which analyzes the use of words and hashtags to recognize patterns, that indicate tourists. This approach includes the creation of a feature vector for each user (e.g.: average time between tweets, time between first and last tweets, most used words), using initial constraints with timestamps to tag an underlying set of tourists and finally classifying tourists using Support Vector Machines. The content of tweets can be transported into analyzable data using term frequency matrices for each user, enriching the feature vectors with these matrices.

Furthermore, the profile of users can indicate whether they are residents of the Netherlands or tourists. It can be checked whether the location of a user, mentioned in his profile, matches with the Netherlands, as more than 80% of all Twitter users provide this type information (Graham, Hale, & Gaffney, 2014).

In order to process the geographical information of tweets, their geotag will be retrieved to enable a visualization on a map. Additionally, the path of movement for each tourist will be generated. Further geographical information about boundaries within the Netherlands will be retrieved from OpenStreetMap[1], enabling the assignment of tweets to areas. Therefore, the following aspects will be addressed within this project:

- The formulation of algorithms to distinguish between residents and tourists
- The creation of visualization approaches
- The analysis and the detection of patterns

The outcome could show how tourists move around in the Netherlands and provide a base for future research on targeted advertisements or services, as well as improving the awareness of tourists.

4 Data

The underlying dataset is described by Tjong Kim Sang et al. (Tjong Kim Sang & van den Bosch, 2013). This dataset provides a collection of tweets starting from 2011, which have either been written in Dutch or are geotagged within the Netherlands or Belgium. For this analysis, geotagged tweets made in 2016 will be considered. Therefore, tweets in different languages can be retrieved. A reduction will be conducted, to

[1] https://www.openstreetmap.de/

only include English tweets, which allows for 1,121,248 tweets. A reduction to English tweets can be explained by tweets generally being dominated by the first spoken language of a country, and English generally dominating other languages in Twitter, as 33.9% of all tweets are posted in English (Magdy, Ghanem, Musleh, & Mokbel, 2007). As *dutch*, the first language in the Netherlands, is being disregarded when analyzing foreign tourists, English remains as the dominating language.

Furthermore, geographical information will be retrieved from OpenStreetMap, regarding outlines of provinces and municipalities. Those can be distinguished by their administrative level. An overview of administrative level and their description can be found in Table 1 (OpenStreetMap, 2018).

Level	Description
2	Boundary of Kingdom of the Netherlands
4	Province boundaries
5	Water board boundary
6	Urban regions in some parts of the Netherlands
7	Collaboration of municipalities
8	Municipality boundaries
9	Districts in Amsterdam and Rotterdam

Table 1: Adminstrative Level (OSM)

Outlines with administrative level 4 and 8 will be collected, in order to assign tweets to provinces and municipalities. The outline of the Kingdom of the Netherlands, illustrated in Figure 1, will be retrieved and used to filter tweets which don't lie within the Netherlands. This enables a disregarding of geotagged tweets made in Belgium. The Overapass API[2] will be used to access OpenStreetMap information.

5 Tourist Detection

While the initial dataset of English tweets in 2016 is huge, not all users can be considered tourists. To distinguish between tourists and residents, classification methods need to be used. Initial constraints, on detecting tourists, have been based on the number of days that have passed between the first and the last tweet of a user. To improve on these detections, two machine learning algorithms were introduced: self training from an initial labeled and unlabeled set of users, and co-training, using user profile and tweet based predictions.

[2] http://www.overpass-api.de/

Figure 1: Outline of the Netherlands, retrieved from OpenStreetMap

5.1 Constraints

To get an insight on how many days Twitter users spend in the Netherlands, the elapsed time between their first and last tweets will be calculated and visualized. Figure 2 shows a histogram, with the x-axis corresponding to the number of days between the first and last tweet, and the y-axis corresponding to the number of twitter users. The height of the graph shows how many users have the given number of days elapsed between their first and last tweet. This histogram shows, that while there are users who spend up to 365 days in the Netherlands, the majority of them spends less than 20 days, with a decreasing trend until 15 days. Due to this, and research conducted by Habib et al. (Habib & Krol, 2017), 15 days were chosen as the initial constraint for tourist detection. Thereby, users will be considered as tourists, who tweeted at least once in the Netherlands, and spent no more than 15 days within the country's territory.

While the constraint detection can reduce the number of unique users needed to be observed, not all of them are useful for the research, as users who tweet only once are not showing any movement within the Netherlands. Using this constraint to select tourists, and discarding

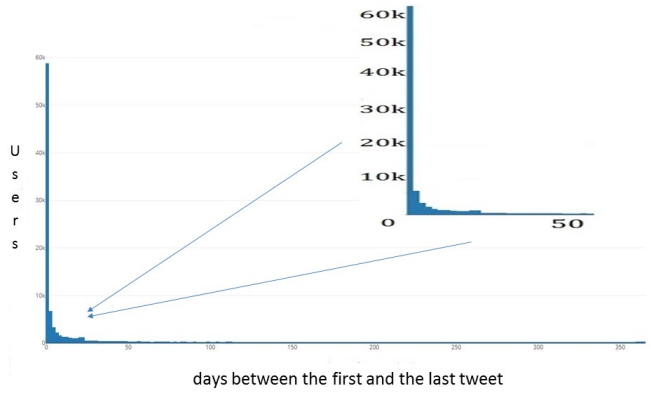


Figure 2: Days between first and last Tweet

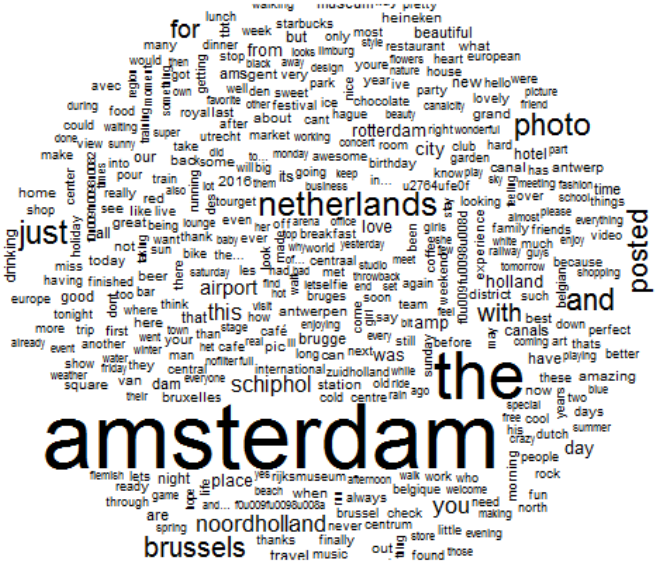
the users who only tweet once, approximately 16,000 moving tourists have been identified. These results carry on into the rest of the research and will be used to conduct results discussed in the paper, as tourist detection was developed parallel to the analysis of tourist movement. Figure 3 illustrates the stepwise reduction of tweets that are considered in this research.

Figure 3: Flow of Data Reduction

5.2 Self Learning

Self learning uses a set of labeled data, where the remaining data is unlabeled. A Support Vector Machine (SVM) classifier is trained on the labeled data and is used to predict the classes of unlabeled data. Once predictions are complete, the algorithm takes the 10 percent most confident predictions, according to the Platt scale (Platt, 1999), and adds these into the training set. The classifier is retrained on an extended set of labeled data, and the steps are repeated until the entire data has been labeled.

The feature vector of a user combines some user information and a term frequency vector of their tweets. Text mining techniques have been used to create a document term frequency matrix from the complete corpus of all tweets. Here, each user is considered as a document, including his respective tweets. After preprocessing and cleaning the data, 98% of sparsity is allowed in the document term matrix. It can be noticed that stop words, such as “to”, “from” and “at” are not removed. While these words usually don’t carry value in text mining, they are important information carriers for tourist detection, as someone might be going *to* Amsterdam, or flying *from* Eindhoven. Each user keeps a number of mentions of the 95 percent least sparse terms in the corpus as their features. Additionally, the following user information is added: time between first and last tweet, country and language profile settings, and whether there are locations mentioned in their profile description. Figure 4 shows the most common words used by the twitter tourists.


amsterdam
the
netherlands
photo
posted
just
and
with
for
airport
this
schiphol
brussels
noordholland
you
rotterdam
holland
canals
place
day
people
night
museum
rijksmuseum
brugge
antwerpen
centraal
station
beautiful
heineken
starbucks
restaurant
hotel
city
canal
antwerp
love
life
time
photos
dam
square
bruxelles
rain
cold
centre
flowers
heart
european
nature
house
party
new
hello
lovely
picture
friend
chocolate
canalcity
grand
hague
beauty
concert
room
club
garden
birthday
awesome
monday
working
market
utrecht
after
about
cant
royal
festival
ice
sweet
den
other
favorite
food
waiting
could
during
sunny
view
done
make
backsome
our
into
pour
train
2016
its
going
keep
business
in...
u2764ufe0f
know
play
sky
meeting
fashion
lover
school
things
everything
almost
please
family
friends
enjoy
video
white
railway
girls
much
because
tomorrow
shopping
district
such
belgian
best
down
perfect
coming
art
thats
playing
better
these
amazing
now
blue
two
days
years
special
free
cool
his
crazy
dutch
summer
morning
rock
fun
north
making
need
welcome
belgique
brussel
check
centrum
never
thanks
finally
travel
music
store
little
evening
found
those
out
big
will
bike
the...
today
miss
drinking
holiday
shop
europe
good
tonight
having
finished
beer
bar
too
saturday
les
had
met
throwback
end
set
again
bruges
soon
team
visit
how
antwerpen
enjoying
come
girl
say
bit
amp
before
sunday
was
next
can
long
international
zuidholland
while
old
ride
ago
station
cold
centre
rain
already
event
another
winter
more
trip
first
went
town
than
stage
café
pic
real
cafe
het
your
man
central
everyone
nofilter
full
show
water
friday
they
weather
van
their
through
game
ready
lets
flemish
ife
and...
beach
when
always
are
spring
thanks
finally
travel
music
out
found
those
evening
store
little
making
north
fun
rock
people
day
summer
dutch
crazy
his
free
cool
years
two
blue
now
these
amazing
have
playing
better
coming
art
thats
perfect
canals
down
best
with
holland
district
such
belgian
posted
and
shopping
because
tomorrow
rainy
white
much
enjoy
video
girls
almost
please
everything
family
friends
looking
things
time
sky
meeting
fashion
lover
school
know
play
hotel
part
canal
has
antwerp
hard

unlabeled set contains 62.061 users. At the end of the learning process, 28.623 tourists have been identified, which leads to the assumption, that if self learning is applied to the whole of the dataset, it would identify approximately 47.000 tourists. However, not only moving users will be identified as tourists. While this is not an issue in tourists detection, it does not provide insight when analyzing the mobility of the tourists, thus the research keeps the originally identified tourists from Section 5.1 for movement visualization and analysis.

5.3 Co-Training

Co-training was introduced by (Blum & Mitchell, 1998) and aims at exploiting information carried in two independent datasets describing the same user. For this to succeed, two separate feature vectors are created for each user: one is purely based on their tweets, and follows the same rules as in self training. The other contains information about user profiles: location, language, named entities, and a text mining approach based on their profile text.

The initial labeled set contains the same users that were used for self learning (1,736 tourists, 5,896 non-tourist). The labeled set is used to train two SVMs, one using the tweet data and the other using the profile data. For sake of simplicity, this paper will refer to the model trained on tweet data as Model A, and the model trained on the profile data as Model B. Table 2 shows the different accuracies of the different SVM models trained in both self learning and co-training.

Model	Self Learning SVM	Model A	Model B
Accuracy	87.1%	86.4%	71.8%

Table 2: Initial Accuracy of SVMs trained on the labeled set

Once the two models are trained, they are used to predict the classes of the unlabeled data set. Following prediction, 10% of the most confident predictions from both models are added to the labeled data set. Confidence is determined using the Platt scale. The process is repeated, until the unlabeled set is empty, and all the instances are labeled.

Co-training identifies 31,267 tourists and 30,794 non tourists from the data set. These are more tourists than self-learning could find, therefore it can be concluded that the profile information enhances the classifier. As with self-learning, co-training does not take into account whether a tourist moves within the Netherlands.

6 Pattern Predictions and Route Recommendation

Examining and mapping tourists leads to revealing travel patterns. These patterns can be exploited, both for knowledge and for recommendations for future tourists who visit the Netherlands and would like to travel an optimal path that is either visited by other tourists as well or corresponds to their travel needs (e.g.: leaving or entering the Netherlands at certain municipalities).

The research examines both a graph based and a data based route suggestion. The data based route suggestion is inspired by Label ranking. Label ranking is one of the 3 tasks of preference learning, which is a subfield in machine learning aiming to learn a predictive preference model from observed preference information. The next section will discuss how the data was brought to fit a label ranking approach, and section 6.3 explains how this data structure was used to find rankings for new tourists.

6.1 Data Preparation

In order to enable route suggestions, routes traversed by tourists will be taken into account. As tweets are marked by their latitude and longitude, a discretization has to be conducted, to group the data and discover patterns in movement between municipalities. Therefore, geotags will be matched to their respective municipality and a movement among those will be analyzed. Movement is indicated by a change in the location of two consecutive tweets. As tourists might tweet multiple, consecutive times from the same municipality, a reduction has to be conducted. This reduction produces the following form:

```
Raw form:
Amsterdam - Amsterdam - Amsterdam - Rotterdam
Reduced form:
Amsterdam - Rotterdam
```

Inevitably, some tourists might return to the same location which they have already visited, however the label ranking algorithm requires unique labels for each municipality. To combat this issue, municipalities appearing multiple times in one path are marked as a separate labels (for instance: Amsterdam, Amsterdam2, Amsterdam3). Table 3 shows the total number of appearances of the most common municipalities in the path of individual users.

There are a total of 361 unique municipalities visited by tourists in the Netherlands. Accounting for returns to the same municipality, a total of 595 labels can be found. The feature matrix for the algorithm is created from these labels, where each instance has

values from 1 to k for each attribute increasing in the order in which they visited municipalities, where k is the total amount of municipalities visited.

Municipality	Max count
Amsterdam	8
Rotterdam	7
Utrecht	7
Zaanstad	7
Haarlemmermeer	6
De Bilt	6
Den Haag	5
Delft	5
Haarlem	5

Table 3: Municipalities visited most times by one tourist

6.2 Graph Traversion

Route suggestions can furthermore be translated into a graph traversion problem, as municipalities can be represented by nodes and a movement among two municipalities as edges. The graph will be populated by placing edges for every change of municipalities encountered in the reduced dataset. In order to model multiple tourists taking the same path, edge weights represent the quantity tourists taking the same path. This will be modeled using a directed graph.

A route can be suggested if arrival and departure municipalities of a tourist, as well as the number municipalities the tourist wants to see during trip, are known. To proceed, every possible path with the requested length *(1 + number of municipalities to see)* will be calculated. Afterwards, a score will be calculated for each path and the path with the highest score will be suggested. As means of assigning scores to paths, each edge will be assigned with a score and edge scores of paths will be summed. Edge scores use an heuristic evaluation based on the relative amount of tourist leaving a municipality for another one.

Figure 5 displays a route suggestion for a tourist arriving and leaving in Maastricht, with the wish to see five other municipalities within the trip. The recommended path can be denoted as: *(Maastricht, Haarlemmermeer, Amsterdam, Drimmelen, Breda, Etten-Leur, Maastricht).*

6.3 K-Nearest Neighbor Prediction

The K-nearest neighbor algorithm is a non parametric pattern recognition algorithm. In this algorithm, all known instances are placed in a 595 dimensional space,

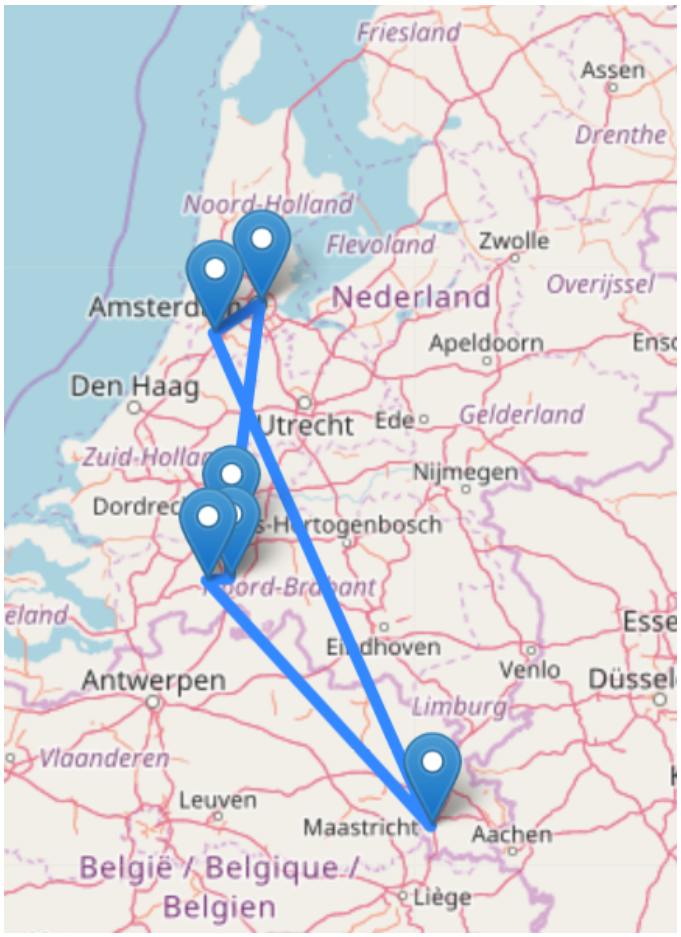


Figure 5: Route from Maastricht - Maastricht with 5 additional cities

where each axis accounts for one label and it's rank. Example instances can be seen on Table 4. NA denotes places the inspected instance (tourist) has not visited. These are modified to 0s before the instance is fed into the algorithm.

userID	Amsterdam	Amsterdam2	Utrecht	Den Haag	Eindhoven	Rotterdam	Maastricht
90367158	4	NA	3	NA	2	NA	1
186374240	1	NA	2	4	NA	3	NA
46448645	1	3	2	NA	NA	NA	NA

Table 4: Example instances for KNN input

Once the instances are mapped onto the 595 dimensional space, a new instance can be created that needs at least 3 inputs: the number of cities a tourists wishes to visit within the Netherlands, the number of neighbors on the graph he wants the predictor to check and a list of initial ordering of municipalities, with at least one municipality on their path. The initial ranking can just be *Amsterdam = 1* to show that the user enters the Netherlands in Amsterdam, or it can be *Amsterdam = 1, Maastricht = 3* to show that the user enters the

Netherlands in Amsterdam, and they want their third visited municipality to be Maastricht. The more initial labels provided, the fewer predictions have to be made.

The algorithm maps the initial ranking onto the 595 dimensional space and checks the k number of closest instances to the ranking. The algorithm finds each missing rank from 1 to n that are not part of the ranking, where n is the total number of municipalities the user would like to visit in the Netherlands. The missing ranks are given to the labels with the most corresponding ranks amongst the neighbors (in case of a tie, the label is decided randomly between the tied labels). A label that already has a rank, can not get a new rank, therefore if it comes up more than once in the prediction, the next most ranked label gets the ranking.

While this prediction works in most cases, there are certain scenarios where it will fail. The first, and most obvious one, is if someone wants to predict for a number of municipalities that has never been inputed in the original dataset (e.g.: predicting for 500 cities fails). The second issue arises due to the numbered labels (e.g.: Amsterdam, Amsterdam2, Amsterdam3). Since the recommender does not understand that these are references to the same municipality, a recommended route can contain multiple of them in a string, or without respect to order. An example recommendation where this goes wrong can be seen on Table 5. In the case of the last tourist, Amsterdam and Amsterdam1 are visited in order, which are different labels, but correspond to the same municipality.

User	Amsterdam	Amsterdam1	Utrecht	Den Haag
Amsterdam start, Amsterdam1 exit	1	4	2	3
Amsterdam start, Den Haag exit	1	0	2	3
Utrecht Start, Den Haag exit	2	3	1	4

Table 5: Recommended cities for 3 different tourists

7 Visualization Approach

Displaying information gets more important with growing amount of data and complexity. In order to not only display data in a tabular form, different kinds of visualization tools can be examined and used. Visualizations enable a intuitive and easy processing, understanding and analysis of complex data. The choice of information visualization tools is therefore of importance to enable a representation of the vast amount of tweets that doesn't cause on overload- and anxiety of information (Wurman, 1989). Shneiderman (Shneiderman, 1996) identified the following key features for information visualization tools:

- Gain an overview on data
- Zoom on specific items
- Use filters to reduce data
- Get detailed information on items
- View relationships between items
- Enable an refinement of actions by keeping a history
- Extract data specified by parameters

In order to comprehend those features, visualizations will be created using the software environment R, specifically the package R Shiny[3]. As means of displaying locations of tweets on a map, Leaflet[4] will be used. Following visualizations approaches can be divided into two subtypes. Firstly, an overall localization of tweet locations will be examined. Secondly, movement of tourists will be examined, by analyzing consecutive tweets of single users.

7.1 Tourist Localization

Tourist localization focuses on visualizing tweets individually on a map according to their geotag. This enables an overview on where users tweet, in order to detect areas with high tweet frequencies. The selection of tweets can be specified and filtered by geographical outlines, given by OpenStreetMap polygons with administration level 4 and 8 (provinces and municipalities). To further distinguish tweeting patterns and tourist behavior, first and lasts tweets of users can be examined separately. This illustrates where tourists arrive in or depart from the Netherlands. Lastly, a filtering by dates is enabled, which also gives the opportunity, to animate tweets over days.

Tourist localization can be visualized using a heatmap (Appendix A) to provide an intuitive overview about popular locations.

7.2 Tourist Movement

In order get an understanding of tourist movement in addition to the location of tweets, tweets can be analyzed per user to extract movement information. This can furthermore be accomplished by visualizing movement on a micro and macro level. The micro level implementation of tourist movement illustrates each path a users take by joining consecutive tweets to one another, which results in paths. Filtering can be conducted by setting constraints on the arrival or leaving of users from the Netherlands. Constraints include limiting time frames of arrival and departure as well as setting specific provinces or municipalities, where users arrived or left the Netherlands. A sample visualization of all routes starting in *Eindhoven* can be found in Appendix B.

[3] https://shiny.rstudio.com/
[4] http://leaflet.js.com/

Conducting an analysis and visualization of tourist movement on macro level groups movement of users on a geographical level of provinces and municipalities. The location of a tweet will be compared with the next tweet of the user, to see whether an area change occurred. By doing so, one can visualize from which area tourist move into a given one, or into what other areas they move, from a given area. This view solely compares two consecutive tweet rather than the entire path a tourist follows. An overview of a macro visualization can be found in Appendix C.

7.3 Features of Visualizations

As noted in the descriptions of visualizations approaches, multiple key features, described by Shneiderman (Shneiderman, 1996), have been considered in the conception. Those include giving an initial overview of the dataset, by displaying the Netherlands in its entirety. Zooming is enabled, to get a closer look on subareas of the map. Furthermore, different filters are provided adjusted for each type of visualization. A display of additional informations is provided, however Shneiderman specifies *"The usual approach is to simply click on an item to get a pop up window with values of each of the attributes"* . This approach isn't followed. Relationships between items can be visualized by using and refining filters, which can be changed freely. A history isn't maintained.

8 Analysis of Patterns

Utilising the visualization approaches described in Section 7, the data can be explored for patterns. Applications of exploration range from getting an overall impression on where tourists are, when tourists tweet as well as insight on how tourists move. Lastly, differences between relevant touristic periods will be contrasted.

8.1 Where are Tourists

The first question that can be addressed when considering tourists within the Netherlands, is where tourists are located. This question can furthermore address the arrival and departure of tourists, as well as where tourists reside in general. A distinction will be made based on municipalities as well as provinces. An extract of the 5 most visited municipalities can be found in Table 6.

A description of the 5 most visited provinces can be found in Table 7. 16,910 tourists have been evaluated for this analysis.

Municipality	Arrival	Departure	Overall
Amsterdam	61%	60%	73%
Haarlemmermeer	10%	11%	17%
Rotterdam	4%	5%	7%
Den Haag	2%	2%	4%
Utrecht	1%	1%	3%

Table 6: Where are Tourists - Municipality

Province	Arrival	Departure	Overall
Noord-Holland	74%	75%	81%
Zuid-Holland	10%	10%	16%
Noord-Brabant	4%	3%	5%
Utrecht	2%	2%	4%
Limburg	2%	2%	3%

Table 7: Where are Tourists - Province

8.2 Where are no Tourists

In accordance to the question of where tourists reside, it can be analyzed whether there are municipalities where no touristic tweets can be found. A distinction can be conducted based on arrivals, departures and overall tweeting behavior. Non-visited areas can furthermore be characterized by the area they occupy (Table 8).

Type	No. of Municipalities	Area in km^2
Arrival	40	3076
Departure	45	3320
Overall	20	1688

Table 8: Empty Municipalities

A visualization of empty municipalities can be found in Appendix D. Analyzing provinces, regarding the question whether they have no touristic activity on Twitter, provides the insight, that tweets are being made in every province.

8.3 When do Tourists Tweet

Following questions on the location of tweets, the posting time of tweets can be considered. Addressing the question, when tourists tweet. An answer will be given based on the weekday of tweets. Figure 6 illustrates the share of tourists tweeting on a given weekday based on arrival, departure and overall behavior.

8.4 Movement - Distances

As tourists are not only tweeting from a single position, their movement can be analyzed. An initial exploration considers the distance between two tweets and the distance tourists cover during their entire stay. Firstly, the

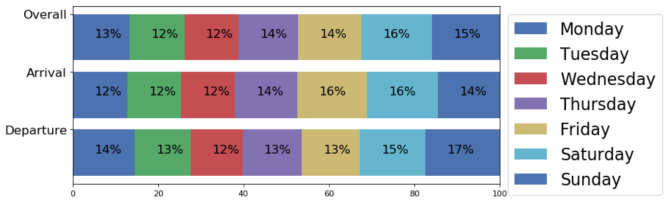


Figure 6: Tweet per weekday

distance between consecutive tweets will be examined. Figure 7 shows a histogram of the corresponding distances. On average, the distance between two consecutive tweets amounts to 7,850m (illustrated by the red, dashed line). The x-axis scale is based on powers of 10. Secondly, the distance a tourists cover will be examined

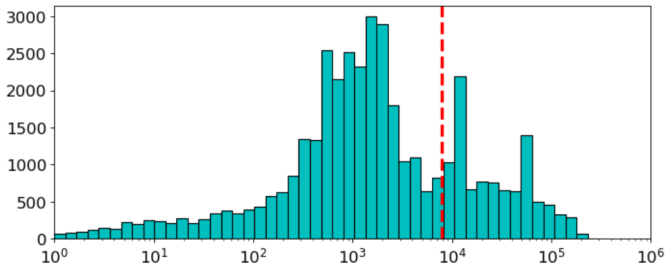


Figure 7: Average distance between 2 tweets

in the histogram in Figure 8. The average traveled distance of tourists amounts to 27,150m. The x-axis scale is set, in accordance to Figure 7, to a scale based on powers of 10.

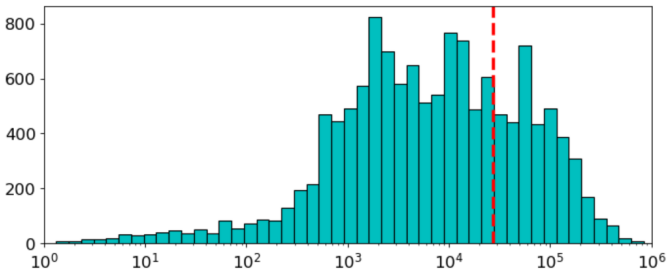


Figure 8: Average distance traveled per User

8.5 Movement - Micro Level

To further deepen insight on tourist movement, tourist paths will be examined as part of the field of movement on a micro level. Every path of users will be considered to represent the Netherlands as a graph based on respective municipalities. The most important touristic cities can be determined by the Pagerank (Page, Brin, Motwani, & Winograd, 1999) of each municipalities. A selection of highest scoring municipalities has been conducted

Municipality	Score
Amsterdam	0.247
Haarlemmermeer	0.111
Rotterdam	0.047
Den Haag	0.036
Zaanstad	0.024
Utrecht	0.021
Edam-Volendam	0.018
Eindhoven	0.012
Lisse	0.012
Haarlem	0.011
Maastricht	0.009

Table 9: Municipality Scoring according to Pagerank

to obtain municipalities distributed across the Netherlands, ranging from Amsterdam to Maastricht. Scores, according to the Pagerank can be seen in Table 9.

A visualization of paths between the most important cities can be seen in Figure 9. The edge width is determined by the amount of tourist taking a path. It can furthermore be denoted that tourists visit 1.47 distinct municipalities on average and move into another municipality 1.6 times. Nonetheless, 11,000 tourists only visited one municipality.

8.6 Movement - Macro Level

Another level of movement that can be analyzed is on a macro level, where two consecutive tweets will be examined for a change in provinces. Aggregating the movement between provinces of all tourists, patterns can be seen in Appendix F. The arrows illustrate a movement of tourists between two provinces. The part of the arrow closest to circle segments, representing provinces, denotes incoming tourists, other arrows denote outgoing tourists. Sizes of circle segments show the amount of tourists that show a movement (either incoming or outgoing) in a province.

On average, tourists visit 1.2 provinces and change provinces 1.27 times. 13,000 tourists visited only one province.

8.7 Touristic Periods

In order to illustrate differences in tourist behavior depending on certain time periods, an average tourist will be compared to touristic periods. Therefore, average number of tweets, active users, arrivals and departures per day will be evaluated and compared. The following touristic periods will be used for comparison:

- Carnival (07.02.2016 - 09.02.2016)