Understanding the Behavioral Differences Between American and German Users: A Data-Driven Study

Chenxi Yang¹, Yang Chen¹, Qingyuan Gong¹, Xinlei He¹, Yu Xiao², Yuhuan Huang³, Xiaoming Fu⁴

¹ School of Computer Science, Fudan University, China
² Department of Communications and Networking, Aalto University, Finland
³ Faculty of European Languages and Cultures, Guangdong University of Foreign Studies, China
⁴ Institute of Computer Science, University of Göttingen, Germany
E-mail: chenyang@fudan.edu.cn, yu.xiao@aalto.fi, huangyuhuan@gdufs.edu.cn, fu@cs.uni-goettingen.de

Abstract: As USA and Germany are the most populous countries in North America and Western Europe, it is meaningful to understand the behavioral differences between American and German users of online social networks. In this work, we conduct a data-driven study based on the Yelp Open Dataset. We demonstrate the behavioral characteristics of both American and German users from different aspects, i.e., social connectivity, review styles and spatiotemporal patterns. In addition, we build a classification model to accurately recognize American and German users according to the behavioral data. Our model achieves a high classification performance with an F1-score of 0.891 and AUC of 0.949.

Key words: Behavioral Difference; Online Social Networks; Yelp; Machine Learning

1 Introduction

Cultural differences have been a core question which sociologists care about. Over the past decades, the cultural differences, the reasons behind these differences and the phenomena that these differences reflect, including collectivism, individualism and social sustainability, have been intensively studied [9, 13, 32]. In these studies, data for understanding users’ culture-related behaviors is obtained using traditional methods, such as questionnaires, video documentation and other personal interview methods.

Compared to rather static online textual data (which stays untouched once published) such as news or web pages, situation-aware interactive information like user reviews and comments provides more daily life related and accessible (typically via text) opinions and thoughts. Nowadays, online social networks (OSNs) [16] have witnessed rapid growth, attracting billions of users around the world. People contribute the profile, social activities and life tracks on OSNs and among these behaviors, deep cultural impacts exist. Krasnova et al. [17] did a survey on Facebook to explore the differences in individual willingness to self-disclosure between American and German users based on a questionnaire to Facebook users. They concluded that American users are more active on Facebook and have higher privacy concerns than Germans. To our knowledge, [17] is the first work which has analyzed users’ online behavior from a cultural perspective. However, as most of the cultural phenomena evolved for many years and developed from generation to generation, the scale of the online survey-based research is, although larger than the survey in the real world, still not large enough to form a cultural impact. Further, data comprehensiveness is of great consequence to cultural analysis. Apart from the answer to the questions in survey or the text posted by users, the
movement pattern and points-of-interest (POIs), which can have a closer relationship with the cultural impact on a user, also matter.

To this end, location-based social networks (LBSNs), such as Yelp [3], Foursquare [6, 33] and Momo [5], which allow users to undertake location-centric activities in addition to social interactions, would offer a viable data source for such cultural studies. These LBSN platforms record the activity data of massive users and provide researchers a great opportunity to compare the human behaviors from both spatiotemporal and social networking perspectives.

USA and Germany, with the largest populations in North America and Western Europe, are two important culture clusters in the world. They have different languages, traditions and geographical conditions, but also share the same Anglo-Saxon origins. Firstly, Germans care more about social stability [20], while Americans tend to become outstanding individuals because of the elite culture [32]. Secondly, in terms of collectivism, Germans prefer risk taking and uncertainty avoidance compared to Americans [34]. However, the differences of German and American users in terms of behavioral patterns are rarely studied. We seek to know whether the behavior of users on LBSNs is consistent with previous cross-cultural research results [8, 17] and if not, which aspects of behavior have changed online.

In this paper, we take data from a representative LBSN, Yelp, as basis for our study. We conduct a data-driven study based on the Yelp Open Dataset*. Yelp can help people discover local businesses, a.k.a., POIs, for example, restaurants, hospitals, or spas. It allows users to publish reviews or conduct check-ins in selected businesses. Yelp users have written more than 142 million reviews by the end of Q3 2017†. Also, Yelp serves more like an “urban guide”, a review platform with location information and category preference, rather than only making friends. This platform partially reflects various aspects of people’s daily life, which makes it possible to infer the entire profile and clearer social engagement of a user. Given its popularity and rich user-generated content, we select Yelp for our user behavior study. Compared to [17], social connectivity, spatiotemporal patterns and writing styles are combined together based on the data from a much larger scale of users in our work. In our study, we pick USA and Germany as the examples for understanding the online behavior from a cultural perspective and make comparisons between this two representative cultural clustering in North America and Western Europe respectively. Our key contributions can be summarised as follows:

- We provide a comprehensive statistical and demographic analysis of American and German users’ behavior based on Yelp Open Dataset, and compare the results in a comprehensive way. We find that American users are more influential on Yelp and their friends are scattering in more cities. Our spatiotemporal analysis shows that compared with American users, German users tend to have a clearer line between daytime and nightlife. According to our analysis of review texts, we also find the evidence that validates the widespread feeling that collectivism is more important for German users while it is individualism for American users.

- In this paper we verify the feasibility of applying big data analysis in the context of cultural behaviors. In particular, we build a classification model, based on our analysis of the users’ online behavior, which is able to accurately detect whether a user is from USA or Germany. With this classification model, we achieve an F1-score of 0.891 and AUC of 0.949 for detecting whether a Yelp user is from USA or Germany, which serves as a strong buttress of our analysis and feature selection. Also, we find that writing style- and social graph-related features are the most distinguishing features to differentiate American and German users.

The rest of the paper is structured as follows: We first introduce Yelp and the dataset used for our study in Section 2. Next in Section 3, we analyze the data for both user groups and businesses on Yelp from a cultural view of USA and Germany and identify which features are strongly related to the cultural diversity. After that, we provide comprehensive evaluations on our classification model using various supervised machine learning algorithms, including the role of importance of different feature sets in Section 4. We review the related work in Section 5, before we conclude the work in Section 6.
2 Background and Dataset

In this section, we will first give an overview of Yelp, and then introduce the open dataset used in this study.

2.1 Background of Yelp

Founded in 2004, Yelp.com has become one of the world’s largest online “urban guide” and business review site [15]. On Yelp, users can write reviews, upload photos, conduct check-ins, and rate the experiences at different types of businesses such as restaurants and hotels. Yelp allows a user to conduct a check-in at a business only when the user is close enough to the business. In addition, users are supposed to give a review of a business several days after the visit. Yelp covers 21 main categories and over 1,200 sub-categories of businesses †. It provides a platform for users to express their preferences over different business categories. Meanwhile, it serves as a social networking platform. Users can make friends with other users who show interests to similar business categories. Together with reviews and check-ins, the information about users’ preferences and friends reflect the user behavior in daily life in an informative way.

2.2 Dataset Description

We study the Yelp Open Dataset, which was used in the Yelp Dataset challenge. The dataset covers over 4,700,000 reviews, 156,000 businesses, 1,100,000 users. Each review contains text and/or rating attributes. The dataset is composed of 11 tables. For each user, we can obtain his/her friends, the year when the user started using Yelp, average number of stars, and some other comprehensive assessment of their reviews and tips. For each business, its location, category, reviews, tips, and check-ins are all available. Regarding each user’s home city, we make the assumption that the user belongs to the city where he or she reviews most, defining the city as the “home city” and getting the country information accordingly. Regarding users’ home countries, USA, Germany, Canada and UK are the four main countries.

3 Data Analysis

In this section, we study the behavior of American and German users based on the Yelp open dataset described in Section 2. Our goal is to better understand the differences and similarities of American and German users in terms of location distribution of friends, social graph characteristics, writing style of reviews, preference for business categories, rating preferences, and the temporal patterns of check-ins. This section is divided into three subsections, i.e., social graph, reviews and check-ins.

3.1 Analysis of the Social Graph

To understand the social behavior of American and German users on Yelp, we use the Stanford Network Analysis Platform (SNAP) [18] for social graph analysis. SNAP is a general purpose network analysis and graph mining library written in C++. Based on SNAP, we analyse some representative network metrics, i.e., degree, clustering coefficient, PageRank and connected components in Yelp’s social network G.

3.1.1 Analysis of the Social Graph as a whole

The Yelp’s friendship network, G, has 8,981,389 nodes and 35,444,850 edges. Fig. 1(a) shows the cumulative distribution function (CDF) of the degrees in G. The degree of a node represents the number of edges connected to the node. A higher degree in G means the user has more friends. There is no nodes with zero degree, which means any user on Yelp has at least one friend. The average degree of G is 7.0. Compared to many other OSNs, 7.0 is a small average degree. As Yelp is not a website dedicated to social-networking, users on Yelp are not chasing for large number of friends. Therefore, the connection on Yelp is looser than in most OSNs.

The clustering coefficient (CC) measures the cliquishness of a typical friendship circle. A higher average CC indicates that it is more likely for nodes to form tightly knit groups. Fig. 1(b) is the CDF of CC of the degrees in G. Over 70% of Yelp users have a CC of zero and the friendship networks’ average CC is 0.055, which demonstrates a weak connection between Yelp users. This is due to the fact that users tend to use Yelp as a guide service rather than a networking site.

PageRank is a metric that quantifies the importance of different nodes in the network [25], which has been applied by the Google search engine to rank the websites. The PageRank value of any node of the network ranges between 0 and 1. A higher PageRank value suggests that the corresponding node is more important in this network. In Fig. 1(c), the CDF of PageRank displays the similar results.

Fig. 1(d) shows the sizes of the top 10 connected components. The connected component is a subgraph

†www.yelp.com/developers/documentationFusion
in which any two nodes are connected to each other by paths, and which is connected to no additional nodes in the supergraph. In network G, there are 18,512 connected components. The largest connected component has 8,938,630 users, covering 99% users in the Yelp Open Dataset. The 2nd largest connected components have 15 nodes and the 3rd to 6th have 14 nodes each. We also calculate the distribution of the shortest path lengths of all node pairs in the largest connect component (LCC), as displayed in Fig. 1(e). The average shortest path length is 4.93 and the average CC of G is 0.055.

We further study the density of the “core” of the Yelp friendship network G. From 0.01% to 10%, we remove some nodes with the highest degrees from the network, and analyse the remaining nodes. As in Fig. 1(f), we group the remaining nodes into three categories, i.e., the LCC, the singletons, and the middle region. We can see that the number of singletons already surpasses that of the nodes still in LCC after we remove 5% of the nodes with the highest degree. Therefore, Yelp’s friendship network is not as strongly connected as other mainstream OSNs, such as Renren [38] and Cyworld [1].

### 3.1.2 Comparison Between American and German Users

Table 1 shows the mean and variance of CC of American and German users’ social graph, which indicates that the mean of CC of American users is slightly higher than that of German users. We also compute the mean and variance of degrees in the American and German users’ social graph and find that for American users, the mean and variance of degree are both larger than those of German users. In other words, some American users tend to make a lot more friends on Yelp.

In Table 1, both the mean and variance of PageRank values of German users are larger than those of American users. According to the PageRank values, we define the top 0.1% users of the whole social graph as “influential users”. We use one set, P, to represent the influential users of the entire graph. We also find that 0.31% of the American users belong to P whereas for German users, that number is 0.18%, which illustrates that both USA and Germany have more influential users than the average level of the entire network of Yelp.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>0.054</td>
<td>0.030</td>
<td>7.0</td>
<td>2379.0</td>
<td>1.180 × 10^{-5}</td>
<td>4.405 × 10^{-8}</td>
</tr>
<tr>
<td>Germany</td>
<td>0.041</td>
<td>0.032</td>
<td>2.0</td>
<td>909.0</td>
<td>2.120 × 10^{-5}</td>
<td>8.360 × 10^{-8}</td>
</tr>
</tbody>
</table>

Table 1 Graph Attributes
Besides, compared to German users, the proportion of influential users is higher in case of American users.

Fig. 2 reflects the location distributions of friends of American and German users. The x-axis represents the number of cities where the users' friends spread among. For the y-axis, the value is the percentage of users who have friends in certain number of cities. As shown in the Fig. 2, the share of German users who only have Yelp friends in one and two cities is greater than that of American users. When it comes to the circumstance that the number of cities is three or higher, the share is about 15% larger in case of American users. We find that German users’ friends are gathered in fewer cities, while the location distribution of American users’ friends is a bit wider than that of German users’.

3.2 Analysis of Reviews

Apart from the analysis of social graph in Section 3.1, we also group the reviews according to the country where the businesses and the user are located. In Section 3.2.1, the text attribute of review serves as an significant part when analysing the character trait. We then give the category preference of American and German users in Section 3.2.2. After that, we present the visit & rating preference in Section 3.2.3.

3.2.1 Writing Style Analysis

Linguistic Inquiry and Word Count (LIWC) [26] is used to fully understand the text of reviews, which has been widely used in computerized text analysis, learning how the words we use in everyday language reveal our thoughts, feelings, personality, and motivations. In our study, as the text of reviews comes from two languages – English and German, we also use the German LIWC2001 Dictionary [35].

As seen in Table 2, the numbers represent the occurrence frequency of that specific kind of words. In the first two columns “Affect” and “Anger”, which represent the affective process with the example words “happy”, “ugly” and “bitter” and contents with “hate”, “kill” and “pissed” [27] appearing having something to do with indulging in anger respectively. The average occurrence of “Affect” and “Anger” of American users are both higher than that of German users. As in [28], American students behave in a more affective way than German students and this conclusion conforms to our results which come from emotions behind writing style.

With regard to the “Tenta” (representing the tentative words) and “Certain” (representing the certainty), American users prefer to write tentative-related words like “maybe”, “perhaps”, “guess” [27], while German users mention certainty-related words like “always” and “never” more frequently. For “Swear”, including words like “fuck”, “damn” and “shit”, there is such a phenomenon that German users are less likely to use swear words about these dimensions when reviewing on Yelp. Words like “buddy”, “neighbor” belonging to the “friends” category appear more frequently in American users’ reviews.

In Table 3, we also find out the differences in occurrence frequency of “I”, “We” and “Leisure” in the six main categories between American and German users. The frequency of writing the reviews with the pronoun of “I” by American users is twice than that of German users. While talking with “We”, the most possible category American users are in is “Nightlife”, whereas it is “Restaurants” for German users. We believe the frequency of the usage of “We” can be positively related to the high prevalence of people going to the particular category of businesses together. Therefore, it is the “Nightlife” and “Restaurants” that serve as the favorite category where American and German users go together respectively.

3.2.2 Preference for Business Categories

Experimentally, our results indicate the distinction of category preference between American and German users. We analyse the distribution of reviews in ten main categories. Fig. 3 displays the category pattern of American and German users, with the y-axis describing the logarithmic coordinates of review percentage of the certain category. For “Food”, “Nightlife” and “Shopping” category, American and German users share similar preferences in general. When it comes to
Table 2 Occurrence Frequency of Different Categories of Words in Reviews

<table>
<thead>
<tr>
<th>Nation</th>
<th>Category</th>
<th>I</th>
<th>We</th>
<th>Leisure</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>Beauty &amp; Spas</td>
<td>7.02</td>
<td>0.53</td>
<td>1.16</td>
</tr>
<tr>
<td></td>
<td>Health Medical</td>
<td>6.85</td>
<td>0.61</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>Home Services</td>
<td>4.89</td>
<td>1.61</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>Nightlife</td>
<td>3.72</td>
<td>1.79</td>
<td>2.68</td>
</tr>
<tr>
<td></td>
<td>Restaurant</td>
<td>4.08</td>
<td>1.49</td>
<td>1.58</td>
</tr>
<tr>
<td></td>
<td>Shopping</td>
<td>5.47</td>
<td>0.88</td>
<td>1.44</td>
</tr>
<tr>
<td>Avg.</td>
<td></td>
<td>5.34</td>
<td>1.15</td>
<td>0.438</td>
</tr>
<tr>
<td>Germany</td>
<td>Beauty &amp; Spas</td>
<td>4.17</td>
<td>0.27</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>Health Medical</td>
<td>3.58</td>
<td>0.31</td>
<td>1.81</td>
</tr>
<tr>
<td></td>
<td>Home Services</td>
<td>2.16</td>
<td>1.05</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>Nightlife</td>
<td>1.73</td>
<td>1.29</td>
<td>1.52</td>
</tr>
<tr>
<td></td>
<td>Restaurant</td>
<td>1.78</td>
<td>1.34</td>
<td>1.42</td>
</tr>
<tr>
<td></td>
<td>Shopping</td>
<td>3.04</td>
<td>0.46</td>
<td>1.15</td>
</tr>
<tr>
<td>Avg.</td>
<td></td>
<td>2.74</td>
<td>0.79</td>
<td>1.44</td>
</tr>
</tbody>
</table>

Table 3 Dimension Values of American and German Users

3.2.3 Visit & Rating Analysis

For the rating section, we also compute the star preference for businesses of American and German users. In the Yelp dataset, the average star of American and German users are 3.73 and 3.78, respectively, with similar variance (1.18 and 1.01). As shown in Fig. 4, the star does not follow a normal distribution and most of the users have a rating of 3-5 in a 5-point scale. German users prefer to grade businesses with a high point but not full. While American users are more likely to give 1 out of 5 or 5 out of 5 compared to German users, who perform a milder rating style.

Fig. 5 displays the CDF of the percentage of users review out of their home city (we call it outer review
number of people visiting other countries are much less than 1%, we analyse the temporal patterns of check-ins at businesses in Fig. 6. The y-axis represents the percentage of check-ins during each hour of a week. According to Fig. 6, American and German users both tend to conduct more check-ins between Monday and Saturday and much less on Sunday. At the same time, the differences of German users’ everyday noon peak (the first peak of a day) and night peak (the second peak of a day) can reach 6%, much more explicit than that of American users, which is only around 0.03%. In addition, given the exact time of everyday, we find that the lunch peak and dinner peak of German users are around 11 a.m and 6 p.m, respectively, while those of American users are around 1 p.m and 10 p.m.

### 3.3 Analysis of Check-ins

As the Yelp Open Dataset only has the check-in information of businesses and the number of people visiting other countries are much less than 1%, we analyse the temporal patterns of check-ins at businesses in Fig. 6. The y-axis represents the percentage of check-ins during each hour of a week. According to Fig. 6, American and German users both tend to conduct more check-ins between Monday and Saturday and much less on Sunday. At the same time, the differences of German users’ everyday noon peak (the first peak of a day) and night peak (the second peak of a day) can reach 6%, much more explicit than that of American users, which is only around 0.03%. In addition, given the exact time of everyday, we find that the lunch peak and dinner peak of German users are around 11 a.m and 6 p.m, respectively, while those of American users are around 1 p.m and 10 p.m.

### 4 Implementation of the CountryClassification Model

After comparing the behavior of social graph, reviews and check-ins between American and German users on Yelp, we build up a broad sense of the behavioral differences between these two groups of users and try to understand the culture influence behind these behavioral differences. From an integrated view, we implement a model based on the features extracted from the analysis results in Section 3 to predict a user’s home country and evaluate the classification performance of our approach and other related proposals. In addition, we also study the importance of each feature in order to look at to what extent these various features can have an effect on American or German users behavior on Yelp.

In this section, we give the implementation details and evaluation process of the classification model. Firstly, we present a brief introduction to the tools and methods used in our study. After that, we describe the training and testing sets. For the last part, we describe the importance of each feature set and make an assessment on the performance during the country classification process. To better implement the machine learning algorithms for the classification model, we adopt XGBoost [4] and Weka [14]. Specifically,
XGBoost is a scalable machine learning system for tree boosting, which is used widely in machine learning contests nowadays. Weka supports a collection of machine learning algorithms for data mining tasks which is implemented in Java. The algorithms in Weka include but not limit to Random Forest (RF), Support Vector Machine (SVM) and C4.5 Decision Tree (J48).

To construct a training and validation dataset, we randomly select 700 American users and 700 German users from the Yelp dataset. We take four representative metrics precision, recall, F1-score and AUC as the standard for evaluating the performance of our classification model. Precision means the fraction of predicted German users who are really German users. Recall represents the fraction of German users who are accurately identified. For F1-score, it is defined as the harmonic mean of precision and recall. AUC is short for “Area Under receiver operating characteristic (ROC) Curve”. Its value is equivalent to the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative example [10]. We adopt several classic machine learning algorithms to train and validate our model using 10-fold cross validation. For each algorithm, we apply grid search to find the “best” parameters set, during which our goal is to achieve a higher F1-score. After the parameter tuning, we randomly select another 350 American users and 350 German users for testing and use the trained model to detect each users’ home country. Table 4 shows our classification performance.

4.1 Evaluating Classification Model as a Whole

As in Fig. 7, we first include all the 25 features as a whole for classification. We divide these features into four sets in Table 5, seven features representing the review counts of 7 main categories in business related feature set, 4 in social graph related, 10 in writing style related and another 4 in visit & rating related feature sets. According to Table 4, we compare several supervised machine learning algorithms including XGBoost, Random Forest, C4.5 Decision Tree, SVM and BayesNet. In particular, we consider SVMp (with polynomial kernel). Using the McNemar’s test [23] to compute the statistical significance, we can see that the prediction performance of nearly every two classifiers are significantly different (p-value<0.005, McNemar’s test). The only exception happens when we consider the Random Forest and XGBoost classifiers, as the difference between them is not that remarkable (p-value>0.2, McNemar’s test). These two classifiers have much better prediction performance than the other ones, implying they both can be used for classifying the cultural belonging in practice. As for our results, with the overall consideration of F1-score and AUC, all the F1-score we achieve from different algorithms are larger than 0.850 and among them, the Random Forest performs the best with an F1-score of 0.891 and AUC value of 0.949. We adopt Random Forest in the following subsection to compare the contributions of different features.

4.2 Evaluating the Contribution of Different Feature Sets

To better understand the importance of different kinds of features in the model, we list the $\chi^2$ (Chi-Square) statistics [36] of top 9 features. As is shown in Fig. 6, we can see that the most discriminating feature is “Pronoun”, which represents words like “I” and “You”. Meanwhile, the features coming from the writing style analysis like “Preps (preposition)”, “Tentat (tentative)” and “Certain (certainty)” are more important than other features. Writing style related feature set plays an important role in distinguishing between the American and German users on Yelp. But social graph related features like “CC”, “Friend_City_Num” and “PageRank” are also of importance, ranking the 5th-7th in Fig. 6. For understanding more details about other feature sets, we also evaluate four feature sets...
Algorithm | Parameter | Precision | Recall | F1-score | AUC
---|---|---|---|---|---
Random Forest | maxDepth=13, numFeatures=5 | 0.893 | 0.891 | 0.891 | 0.949
XGBoost | learning_rate=0.01, min_child_weight=1, max_depth=4, gamma=0.0, subsample=0.95, lambda=1, alpha=0, colsample_bytree=0.75, boost=gbtree, objective=binary:logistic | 0.891 | 0.890 | 0.890 | 0.901
Decision Tree (J48) | confidenceFactor C=0.2, Instance/Leaf M=4 | 0.878 | 0.878 | 0.878 | 0.899
SVMp | Degree=3, Cost parameter=20.0 | 0.853 | 0.853 | 0.853 | 0.853
BayesNet | default | 0.862 | 0.862 | 0.862 | 0.923

Table 4 Comparison of Different Supervised Machine Learning Algorithms for the Classification Model

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business</td>
<td>Restaurants and Food&lt;br&gt;Nightlife&lt;br&gt;Event Planning &amp; Services&lt;br&gt;Hotel &amp; Travel&lt;br&gt;Art &amp; Entertainment&lt;br&gt;Beauty &amp; Spas&lt;br&gt;Health &amp; Medical</td>
</tr>
<tr>
<td>Social Graph</td>
<td>Number of cities having friends&lt;br&gt;Clustering coefficient&lt;br&gt;Degree&lt;br&gt;PageRank</td>
</tr>
<tr>
<td>Writing Style</td>
<td>Number of words per sentence&lt;br&gt;Frequency of occurrence of preposition&lt;br&gt;Frequency of occurrence of pronoun&lt;br&gt;Frequency of occurrence of “Anger”&lt;br&gt;Frequency of occurrence of “Leisure”&lt;br&gt;Frequency of occurrence of “Sad”&lt;br&gt;Frequency of occurrence of tentative words like “maybe”, “perhaps”&lt;br&gt;Frequency of occurrence of certain words like “always”, “never”&lt;br&gt;Frequency of occurrence of “Friends”&lt;br&gt;Frequency of occurrence of swear words</td>
</tr>
<tr>
<td>Visit &amp; Rating</td>
<td>Number of reviews&lt;br&gt;Number of visited cities&lt;br&gt;Percentage of visit out of home city&lt;br&gt;Average star</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>χ²</th>
<th>Feature</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>969.876</td>
<td>Pronoun</td>
<td>Writing Style</td>
</tr>
<tr>
<td>2</td>
<td>650.939</td>
<td>Preps</td>
<td>Writing Style</td>
</tr>
<tr>
<td>3</td>
<td>366.716</td>
<td>Tentat</td>
<td>Writing Style</td>
</tr>
<tr>
<td>4</td>
<td>268.432</td>
<td>Certain</td>
<td>Writing Style</td>
</tr>
<tr>
<td>5</td>
<td>199.665</td>
<td>CC</td>
<td>Social Graph</td>
</tr>
<tr>
<td>6</td>
<td>99.615</td>
<td>Friend_City_Num</td>
<td>Social Graph</td>
</tr>
<tr>
<td>7</td>
<td>85.471</td>
<td>PageRank</td>
<td>Social Graph</td>
</tr>
<tr>
<td>8</td>
<td>73.282</td>
<td>Swear</td>
<td>Writing Style</td>
</tr>
<tr>
<td>9</td>
<td>60.701</td>
<td>Beauty &amp; Spas</td>
<td>Business</td>
</tr>
</tbody>
</table>

Table 5 Subsets of Features of the Classification Model

Table 6 Feature Importance: χ² Analysis

independently and compare their performance. Results are shown in Table 7. With the 10 writing style related features, we achieve an F1-score of 0.878 and AUC of 0.937. For the social graph related feature set,
<table>
<thead>
<tr>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Writing Style</td>
<td>0.879</td>
<td>0.878</td>
<td>0.878</td>
<td>0.937</td>
</tr>
<tr>
<td>Social Graph</td>
<td>0.741</td>
<td>0.741</td>
<td>0.741</td>
<td>0.823</td>
</tr>
<tr>
<td>Business</td>
<td>0.623</td>
<td>0.612</td>
<td>0.617</td>
<td>0.660</td>
</tr>
<tr>
<td>Visit &amp; Rating</td>
<td>0.602</td>
<td>0.603</td>
<td>0.601</td>
<td>0.619</td>
</tr>
</tbody>
</table>

Table 7 Contribution Analysis of Each Feature Subset in Classification Model

F1-score is 0.741 and AUC is 0.823. Then for the other two feature sets, F1-score is 0.617 and 0.601 while AUC is 0.660 and 0.619 of business related features and visit & rating related features, respectively.

Corresponding to the Chi-Square statistical analysis, the differences in writing style and social graph are the two most distinguishing attributes between users from USA and Germany on Yelp. Furthermore, preferences for business and visit & rating have a relatively slight effect on classification.

5 Related Work

Combining the features of social networks with geographic information sharing, LBSNs are becoming increasingly popular. Some recent works focused on exploiting online social interactions among individuals for explaining social phenomena. For example, Topa et al. [31] generated a stationary distribution of unemployment which exhibited positive spatial correlations, and Wal et al. [30] applied the social network analysis in economic geography. For the friendship distribution and user mobility, Liben-Nowell et al. [19] introduced a model capturing user behaviors in real-world social networks and found that the probability of befriending a particular person and the number of closer people are inversely proportional. Cho et al. [7] discovered that compared to short-ranged travels, long-distance travels are more likely influenced by social network ties. Those studies provided strong evidence that data on LBSNs can be utilized to analyse groups of users behavior. However, to our best knowledge, no studies have leveraged LBSNs for understanding cultural differences between the behaviors of users from different countries.

In the past decade, Yelp has become a worldwide online business review site, which records millions of reviews and business preferences of users from different countries. With its Open Dataset, lots of researchers did studies on Yelp. Byers et al. [3] studied the correlation between the groupon behavior of a business and the users rating distribution to this business. On the Yelp Open Dataset, fake reviews or even the malicious reviewers were filtered out by analyzing the text of reviews [21, 24, 37]. Moreover, [2, 15] leveraged the rating and category or business preference extracted from the keywords or sentences in text, for analysing the users appetite so as to understand users’ feedback. Furthermore, text and rating were also used together for understanding users’ feedback [22]. In this work, we combine the text, rating and reviews on Yelp, which forms a more comprehensive understanding of user behaviors from two cultural clustering.

Towards understanding user behaviors from a cross-cultural perspective, a preliminary work [29] figured out that the different cultural background has impacts at the individual level on IT acceptance. To our knowledge, [17] is the first work which applied cultural analysis methods while studying the self-disclosure behavior on OSNs. They did a survey on Facebook and explored the differences in individual willingness to self-disclosure between American and German users. [11] also analysed the school pupils’ daily habits in Germany and China. Unfortunately, both studies [11, 17] relied on data collected from online surveys, the scale of which was quite limited. Garcia-Gavilanes et al. [12] considered the text-based “big data” on Twitter for finding whether there is a strong relationship between users’ behavior on Twitter and the traditional cultural theories. However, even this work [12] explored a large scale of users, aforementioned works [11, 12, 17] lacked of richer features in their data, which are largely based on either answers to a questionnaire or the “free-text” on Twitter. In addition, for [12] particularly, they cared about the relationship between the user’s online behavior and real culture phenomenon, but told little about the differences of users’ behavior between certain culture clustering, i.e. cross-cultural behavior. To narrow the gap between comprehensive data and online cross-cultural behavior, our work combines social connectivity, spatiotemporal patterns and writing styles of users on Yelp to understand the differences and similarities between American and German users.

We build on these past works to study the feasibility of using review-based website to analyse cultural
differences between certain cultural clusterings.

6 Conclusion

By referring to the Yelp Open Dataset, we use the behavioral information of massive users to explore the differences between American and German users. We find that the major differences are category preference, mobility pattern, friendship distribution, rating preference and writing style. We utilize the results to extract several human behavior patterns and generalize their features. In addition, we build a classification model to detect where a certain user comes from based on the features we extract. With this model, we validate our analysis results and also gain better understanding of the importance of various feature sets in forming a human behavior pattern on Yelp.

There are still a lot to explore about the cultural causes of user online behavior. Expanding the variety of social platforms, for example, to other LBSNs like Foursquare, would be an important future work for us to study the cultural differences, in combination with the study here using Yelp. Meanwhile, doing an offline user study is also on our schedule. We aim to build an overall behavior pattern of cultural consequences which can apply to people with different cultural backgrounds.

References


Chenxi Yang is an undergraduate student within the School of Computer Science at Fudan University, China. She has been a student research assistant in the Mobile Systems and Networking (MSN) group since 2016. She visited Peking University (in 2017) as a research intern. Her research interests include massive data analytics and machine learning.

Yang Chen is an Associate Professor within the School of Computer Science at Fudan University, China. Before that, he was a postdoctoral associate at the Department of Computer Science, Duke University, USA, where he served as Senior Personnel in the NSF MobilityFirst project. From September 2009 to April 2011, he has been a research associate and the deputy head of Computer Networks Group, Institute of Computer Science, University of Göttingen, Germany. He received his B.S. and Ph.D. degrees from Department of Electronic Engineering, Tsinghua University in 2004 and 2009, respectively. He visited Stanford University (in 2007) and Microsoft Research Asia (2006-2008) as a visiting student. His research interests include online social networks, Internet architecture and mobile computing. He is serving as an Editorial Board Member of the Transactions on Emerging Telecommunications Technologies (ETT) and IEEE Access. He is a senior member of the IEEE.

Qingyuan Gong received her B.S. degree in Computer Science from Shandong Normal University in 2012. She is now a PhD candidate in Computer Science at Fudan University, China. Her research interests include network security, user behavior analysis and modeling, and distributed storage systems. She published referred papers in ACM Transactions on the Web, IEEE Communications Magazine and IEEE ICPP. She has been a visiting student at the University of Göttingen in 2015.

Xinlei He received his B.S. degree in Computer Science from Fudan University in 2017. He is now a master student in Computer Science at Fudan University. His research interests include social network analytics and data mining.

Yu Xiao received her doctoral degree (with distinction) in computer science from Aalto University in January 2012. Before that, she got her Master and Bachelor degrees in computer science and technology from Beijing University of Posts and Telecommunications, China. She is currently an assistant professor in Department of Communications and Networking, Aalto University where she leads the mobile cloud computing group. Her research interests include edge computing, mobile crowdsensing, and energy-efficient wireless networking. Her work has received 3 best paper awards from IEEE/ACM conferences. She is also a recipient of the 3-year postdoc grant from Academy of Finland.

Yuhuan Huang is an assistant professor within the Faculty of European Languages and Cultures at Guangdong University of Foreign Studies. She received her doctoral degree (with distinction) in Intercultural German Studies from the University of Göttingen, Germany in May 2017. Before that, she received her Master and Bachelor degrees in German Studies from Beijing Foreign Studies University, China. Her research interests include German language and culture as well as intercultural communication.

Xiaoming Fu is a full professor of computer science at the University of Göttingen. He received his Ph.D. from Tsinghua University in 2000. He was then a research staff at TU Berlin before moving to Göttingen in 2002, where he has been a professor and head of computer networks group since 2007. His research interests lie in networked systems and applications, including mobile and cloud computing, social networks and big data analysis. He currently serves on the editorial boards of IEEE Communications Magazine, IEEE Transactions on Network and Service Management, and Computer Communications.