ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 11, November 2017

A Survey on User Behavior Clustering in Multimedia Social Networks

[1] Swetha Koduri, [2] T. Satya kiranmai
[1][2]Assistant Professor
[1] Information Technology,Malla Reddy College of Engineering and Technology,Maisammaguda, Dhulapally, Telangana.
[2] Computer science and engineering, CMR College of Engineering and Technology, Kandlakoya, Medchal, Telangana.

***Abstract*— This Survey investigates how quantitative user data, extracted from server logs, and clustering algorithms can be used to model and understand user-behavior. The Survey also investigates how the results compare to the more traditional method of qualitative user-behavior analysis through observations. The results show that clustering of all user data, only a small subset of users increases the reliability of findings. However, the quantitative method has a risk of missing important insights that can only be discovered through observation of the user. The conclusion drawn in this survey is that a combination of both is necessary to truly understand the user-behavior.**

Keywords— Multimedia , User-behavior, quantitative;

1. INTRODUCTION

Understanding user-behavior is important in many contexts. This includes user experience design, software development and marketing. The traditional method for understanding user-behavior is based on qualitative data, which means that in order to learn how users of a certain service use the product the user has to be observed and the observations has to be analyzed. There are many tools available in order to make the process easier, for example for the analysis of a user interacting with a computer program, screen recording software means that no observer has to be in the same room. Still, the analysis is based on observations of specific users and since the analysis of each user can take some time, the whole process forces the conductor of the tests to make a trade-off between sample-size and time. With every extra user in the sample, the time and therefore cost of the analysis becomes bigger. And even if hundreds of users are observed using this method, they will still only represent a small section of the user communities of hundreds of thousands, or even millions, that many online services have today. This means that there is a large statistical risk of choosing a non-representative sample group, which can lead to user-behavior patterns being under or over represented in the group. This can lead to incorrect conclusions which can be very difficult to verify or dismiss later on. The intent of this Survey is to present an alternative method for understanding user-behavior.

Objective:-

The objective of this thesis is to understand how useful a quantitative methodology, based on data-mining and clustering algorithms, is for determining user-behavior, compared to the more traditional qualitative methodology that is commonly used today.

Purpose:-

The purpose of this survey was to expand the knowledge regarding how to identify user-behavior in general, and identifying user-behavior on collaborative Multimedia Social Networks.

Limitation:-

The focus of this study is on the identification of user archetypes, which means that users were classified into one of a number of groups based on archetypical behavior. The intent of the study was to draw general conclusions regarding the viability of

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 11, November 2017

the method by using it in a specific case. The archetypes that will be identified may therefore look very different when the method is used in other contexts.

2. LITERATURE SURVEY

a. Characterizing User Behavior in Multimedia Social Networks(MSN):-

In this study, the authors looked at user data on four MSNs collected through a Multimedia Social Networks aggregator. The aggregator lets users access and interact with multiple MSNs through a single website. The study looked at many different aspects of user behavior, such as session durations, number of logins and transitions between different activities.

These activities were categorized into six categories:

1. Search

2. Messages (Private messages)

3. Videos (Browsing and viewing)

4. Photos (Browsing and viewing)

5. Profile & Friends (Browsing your, your friends' or your friends' friends' pro- files)

6. Communities (Browsing and posting in communities)

b. Identifying User Behavior in Multimedia Social Networks:

In this study, the authors used data mining and statistical methods in order to analyze the users of YouTube. For each user a vector of nine parameters was created, based on data gathered by observing network traffic.

The nine parameters analyzed were:

1. Number of uploads

2. Number of watches

3. Number of channel views (channels visited)

4. System join date

5. Age (Time from join to latest login)

6. Clustering coefficient

7. Reciprocity (Probability of mutual subscription)

8. Out-degree (Number of subscriptions)

9. In-degree (Number of subscribers)

The authors used K-means to cluster the data, and in order to find the optimal number of clusters, they used two different methods. The first method, minimizing Coefficient Variation, resulted in no number of clusters being more optimal than another. The second method, iterating over number of clusters K until the distance between two centroids was below a certain threshold, gave the result 5 clusters for threshold.

The authors discuss the validity of choosing an arbitrary threshold, but suggest no other ways of doing so. The five identified clusters were defined by the authors as:

1. Small Community Members (6%)
2. Content Producers (23%)
3. Content Consumers (13%)
4. Producers & Consumers (48%)
5. Other (10%)

c. Using Cluster Analysis in Persona Development:

In this survey, the authors describe a case study done, in which they applied cluster analysis to create personas for a travel. The purpose of the

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 11, November 2017

study was to understand whether a quantitative method could be used instead of the conventional qualitative method.

The authors created a survey with 45 dimensions. 27 of the dimensions were used in the clustering process, as they were focused on behavior. The rest were used as additional data for the persona creation process, and contained only demographic data.

The clustering dimensions were then used as parameters to create two clusters with the help of the K-means clustering algorithm. Two personas were created, based on the respondents closest to the average in the two clusters. The remaining 18 dimensions of the survey were used to make the personas more complete. The conclusion drawn by the authors is that personas can be created with this method, but since no qualitative methods were used it is not possible to compare the quality between the two methods.

3. CLUSTERING

K-Means is one of the oldest methods for clustering. The purpose of the algorithm is to find a way to partition a set of data that minimizes the sum of squared errors over all the clusters. This problem is a proven NP-hard problem, so the algorithm begins with an initial guess of cluster midpoints and then iteratively tries to improve the partitions. This means that the algorithm will only find a local optimum, and the global optimum is not guaranteed.

One of the biggest challenges when categorizing data is determining the number of categories. This problem has been researched many times. The authors explain the methodology of the elbow method. The method, which can be used both for finding the number of principal components and the number of clusters in a set of data, consists of plotting a downward sloping graph and identifying the elbow; the place where the graph goes from having a vertical slope to a horizontal slope. The more pronounced the elbow is, the more certain the number is.

A gradual decline without a pronounced elbow means that there is no clear optimal number. When finding the optimal number of clusters, the sum of squares within the clusters should be plotted against the number of clusters. The elbow will then represent the point where adding more clusters does not significantly affect the total sum of squares. The method consists of comparing the compactness of the clustering, with the same clusters applied to a uniformly random generated dataset. The number of clusters which result in this biggest gap in compactness between the real data set and the generated data set is considered the optimal number.

Algorithm explain: When all the data had been extracted and stored it was analyzed using the statistical computing language R2 . K-means was used for the clustering . The implementation used is the one in the R library Stats. The classifications were done with 25 random starting points and a maximum iteration of 1000.Sum of squares was used for determining the number of clusters, together with the elbow method, but the results were inconclusive. This suggests that the data might be too complex for a distinct classification algorithm. However, since the purpose of the classification in this research was to expose underlying patterns, rather than to create groups of users, a perfect clustering was not necessary. In order to produce clusters of users, an approximate elbow was used.

4. CONCLUSION

The purpose of this thesis was to answer the question whether quantitative methods can be used to understand user-behavior. The advantage of using

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 11, November 2017

quantitative methods is indisputable, the statistical certainty of doing the analysis on all users means that findings can be used with confidence. When using qualitative methods, it is inevitable that only a subset of users will be analyzed. This will lead to some degree of uncertainty whether the examined set of users is representative of the entire user base. However, there are also flaws in the quantitative method. Mainly, two things are difficult to detect when using clustering. The statistical method may show that a user does not upload content to the service. Two very different problems with very different solutions, but they would be difficult to differentiate between using a statistical method. The surrounding behavior relates to behavior that is invisible to the quantitative method. This could include, sharing content on the service through a third-party communication service like Facebook, or going to another website like YouTube to find information. Since the method described in this survey only looks at activity on the service, it is impossible to say whether someone switches to YouTube because they wanted to find more information, or because they got bored. The conclusion drawn in this thesis is that currently, the best method for understanding user-behavior is a combination of both qualitative and quantitative methods.

5. FUTURE RESEARCH

First, the viability of using quantitative method to create personas and understanding user-behavior should be further analyzed. Methods for cross-service data collection could greatly increase the depth of analysis possible. Second, the archetypes developed using the quantitative method were likely generic enough to fit many other MSNs.

6. REFERENCES

[1] David A. Siegel. The mystique of numbers: belief in quantitative approaches to segmentation and persona development. In SIGCHI Conference on Human Factors in Computing Systems, pages 4721–4732. ACM, 2010.

[2] Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida. Characterizing user behavior in online social networks. In Proceedings of the 9th ACM SIGCOMM Internet measurement conference, pages 49–62. ACM, 2009.

[3] Marcelo Maia, Jussara Almeida, and Virgílio Almeida. Identifying user behavior in online social networks. In Proceedings of the 1st workshop on Social network systems, pages 1–6. ACM, 2008.

[4] Tom Wilson. Data mining user behavior. CMG MeasureIT, 2010.

[5] Francis T. O'Donovan, Connie Fournelle, Steve Gaffigan, Oliver Brdiczka, Jianqiang Shen, Juan Liu, and Kendra E. Moore. Characterizing user behavior and information propagation on a social multimedia network. In Multimedia and Expo Workshops (ICMEW), pages 1–6. IEEE, 2013.

[6] Moira Burke, Robert Kraut, and Cameron Marlow. Social capital on Facebook: Differentiating uses and users. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 571–580. ACM, 2011.

[7] Myra Spiliopoulou, Lukas C. Faulstich, and Karsten Winkler. A data miner analyzing the navigational behaviour of web users. In Proceedings of the Workshop on Machine Learning in User Modelling of the ACAI, Greece, volume 7, 1999.

[8] Jennifer Jen McGinn and Nalini Kotamraju. Data-driven persona development. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 1521–1524. ACM, 2008.

[9] Jon Brickey, Steven Walczak, and Tony Burgess. A Comparative Analysis of Persona Clustering Methods. In AMCIS, page 217, 2010.

[10] Nan Tu, Xiao Dong, P. Rau, and Tao Zhang. Using cluster analysis in persona development. In SCMIS, 2010.

[11] Lieve Laporte, Karin Slegers, and Dirk De Grooff. Using correspondence analysis to monitor the persona segmentation process. In Nordic CHI: Making Sense Through Design, pages 265–274. ACM, 2012.

[12] Rashmi Sinha. Persona development for information-rich domains. In SIGCHI Conference on Human factors in computing systems, pages 830–831. ACM, 2003.

[13] Cooper. Getting from research to personas: harnessing the power of data | Cooper Journal.

[14] Seung Ha, Hong Jung, and Yong Oh. Method to analyze user behavior in home environment. Personal and Ubiquitous Computing, 10(2-3):110–121, 2006.

[15] David R. Thomas. A general inductive approach for analyzing qualitative evaluation data. American journal of evaluation, 27(2):237–246, 2006.

[16] Anil K. Jain. Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8):651–666, 2010.

[17] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2):411–423, 2001.

[18] Siddheswar Ray and Rose H. Turi. Determination of number of clusters in kmeans clustering and application in colour image segmentation. In Proceedings of the 4th ICAPR, pages 137–143. IEEE, 1999.

[19] Dan Pelleg, Andrew W. Moore, and others. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In ICML, volume 1, 2000.

[20] Mu Zhu and Ali Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. Computational Statistics & Data Analysis, 51(2):918–930, 2006.