Media and Communication Research (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/mediacr.2024.050225
ISSN 2523-2584 Vol. 5 Num. 2

Clustering-based Creator Correlation Analysis of Little Red Book

Wenxuan Chen[1,a,*]

[1]United World College of Singapore, Singapore
[a]15618517686@163.com
[*]Corresponding author

Keywords: Clustering, Little Red Book, Random forest

Abstract: The proliferation of social commerce platforms, exemplified by Little Red Book, has created unprecedented opportunities for analyzing consumer behavior through user-generated content. This paper introduces a pioneering clustering analysis model tailored to the distinct environment of Little Red Book. Our model integrates advanced data mining techniques with natural language processing to systematically categorize user posts, reviews, and interactions, aiming to decipher complex consumer behavior patterns. By employing this model, we are able to identify distinct consumer clusters based on preferences, engagement levels, and sentiment towards products and brands. This segmentation enables a deeper understanding of market trends, user needs, and potential areas for product innovation. Our methodology is validated through a series of experiments, demonstrating the model's effectiveness in providing actionable insights for marketers, product developers, and platform managers. The findings underscore the potential of targeted clustering analysis in enhancing strategic decision-making in the digital marketplace. This study not only contributes a novel tool for academic and practical applications within the realm of social commerce analytics but also sets a foundation for future research in consumer behavior analysis on digital platforms.

1. Introduction

Clustering analysis is a powerful tool in the realm of data science, offering a way to identify natural groupings within large datasets based on shared characteristics. In this context, we presents a novel approach to leveraging clustering techniques specifically tailored for the social media platform Little Red Book [1]. Little Red Book is a popular social commerce platform in China that combines lifestyle sharing with product reviews, making it a rich source of user-generated content data. This article introduces a comprehensive model designed to analyze and interpret the vast amounts of unstructured data generated on Little Red Book, aiming to uncover patterns and trends that can inform marketing strategies, product development, and user engagement initiatives. By focusing on the unique characteristics of Little Red Book's data, the model seeks to provide valuable insights into consumer preferences and behaviors, offering a strategic advantage in the competitive landscape of digital commerce [2].

2. Literature review

In current days, many people use social media during their free time to relax by scrolling though recommended posts and videos that are published by other users. As social media applications evolve, more and more users begin to share their ideas online and it becomes harder to determine what the users will receive on their recommendation page. In this case, many companies began to use special prediction methods to generate the most relatable posts to each target user, which they take factors such as users' interest and interaction with other users into consideration and recommend contents that seems most suitable for them [3].

Chia-Chuan Hung analyzed the role of tagging on the social interactions between users on social media [4]. To achieve this, the team developed a tag-to-tag matrix that records the relationship between a pair of user tags and content tags and presents an index indicating the relevance between the tags. However, through using the tag-to-tag matrix, the team only managed to obtain a very low precision and didn't manage to come to a certain conclusion.

In the following years, many other researchers have conducted research on social media analysis and content mining through different approaches. In Chung-Hong Lee's (2012) research on evaluating social media contents, a density-based clustering method is adopted to due to the amount of irrelevant information's contained in social media posts. Using density-based clustering methods, the irrelevant notices will be treated as outliers which minimizes their impacts on the model. Tang J and Liu H pproached in investigating the ways to model the relationship among data instances and the use of the model in feature selection. By proposing the concept of pseudo-class labels which is later used to guide the extraction of linked information and attribute-value information in posts, they developed a novel framework Linked Unsupervised Feature Selection (LUFS) that predicts relevance between posts. After comparing the results with three other unsupervised feature selection algorithms: Unsupervised Discriminative Feature Selection (Y. Yang, 2011) and Laplacian Score [5], it has been shown that the LUFS method is able to perform significantly more accurate than the other three existing baseline methods.

A comparative study of supervised models was later conducted in approach to determine the most suitable selection method for social media news feed using the features of the users and post (Belkacem, S., Boussaid, O. and Boukhalfa, K, 2020). After identifying 16 related work features that influences the relevance of a news feed and comparing them to the news feed relevance, it had been found that Gradient Boosting (Natekin, A. and Knoll, A., 2021) and Random Forest (Biau, G. and Scornet, E., 2016) has shown much more accurate predictions in user preferences. In a sentiment analysis of social media text (Rahman, H, 2023), a multi-tier model comprising three models is created to perform supervised learning. In this model, data and texts are classified into different levels of positive and negative classes, with each classifier being trained using Naïve Bayes (Rish, I., 2001), Decision Tree (Song, Y.Y. and Ying, L.U., 2015), and SVM (Joachims, T., 1998). These classes are later used to predict sentiments of text data, such as labels of unseen movie reviews [6].

In these existing research, many used clustering methods to model the relationship between users and posts, and some other used different supervised learning methods to predict the relevance of social media recommendations. As there is no existing data of post relevance in the media platform that we will be studying on, it will be impossible to perform a supervised learning method, but we can still approach through clustering methods using data of different features on existing media posts. In the sentiment analysis, the model is created to analyse the sentiments of text data. Although it is not suitable to apply the same model to analyse the relationship between users and posts, it is possible to implement the idea of classifying data and texts onto the analysis. Therefore, our plan is to identify the determinants of social media relevance on users and use this information to create a model through clustering, which predicts the possible methods that are used in the media platform that we are focusing on [7].

3. Research Methods & Results

3.1. Longitudinal Data of Red Book

Red book longitudinal data research is public data we obtained based on python. The data content involves the blog posts of the top bloggers of each topic in red book from 2023.1.1 to 2024.1.1, the number of fans, publishing time, blog post topics, etc. There are about 10K pieces of data in total.

3.2. Data Management and Cluster Analysis

As Table 1 mentioned, data management is divided into two steps: first, we cluster based on the likes of multiple blogs and videos from bloggers, and use Python's graph theory analysis of clustering with graph 0.8.3. Secondly, use R version 4.0.5 and ggplot for further analysis and visualization.

Table 1: Network Properties That Were Calculated for Clusters and Nodes

Variable	Object	Description
Node degree	Node	Number of links
Past node growth	Node	Number of links gained over the past 3 month
Future node growth	Node	Number of links gained over the next 3 month
Closeness	Node	$(n-1)/(\Sigma_i^n p_i)$, where n is the number of nodes and p_i is the shortest path from the node of interest to node i
Betweenness	Node	Number of shortest paths between each pair of
Cluster size	Cluster	Number of nodes in the cluster
Past cluster growth	Cluster	Number of nodes gained over the past 3 month
Future cluster growth	Cluster	Number of nodes gained over the next 3 month
Density	Cluster	m/(n×(n-1)/2), where m is the total number
Transitivity	Cluster	Probability of 2 neighbors of the same node
Median degree	Cluster	Median of all node degrees in the cluster
Median distance	Cluster	Median Tamura-Nei 93 distance of all the links
Median closeness	Cluster	Median of the node closenesses
Median betweenness	Cluster	Median of the node betweennesses

3.3. Cluster- and Node-Level Growth Modeling

We used Poisson regression to model the number of nodes acquired in each cluster from 2023.1.1 to 2024 and assess factors associated with cluster growth. One considered factor was the past cluster growth (defined as the change in cluster size from 2011 to 2014). We used logistic regression to model the acquisition of new links of individual nodes within the first 3 years of being enrolled in the cohort (with a binary outcome variable). We included variables that have been found to be predictive of cluster growth or clustering in similar work, such as favourite count and type of blog [8].

3.4. Modelling

We use Poisson regression to simulate the amount of data each cluster obtains from January 2023 to August 2023 and evaluate factors associated with cluster growth. Factors considered include past cluster growth (defined as the change in cluster size from January 2023 to August 2023). We used logistic regression to model the occurrence of individual data points acquiring new connections within the first 3 years of joining the study (binary outcome variable). Included are some variables that have been found to predict cluster growth or clustering in similar work, such as age, gender, activity, user engagement, and content preferences.

3.5. Cross-validation

To predict whether nodes will acquire new links in the next 3 months, we used a 10 fold cross validation method to avoid overfitting and compared logistic regression with random forest classification based on several variable subsets. Train logistic regression models and random forest models on the same training data for each set of predictive factors [9].

3.6. Results

Analyzing a total of 13299 sequences with the distance-based clustering algorithm yielded a total of 998 clusters that were highly robust over the observed time frame, making it possible to assess the dynamics of the clusters and their constituent nodes over the 12 month-long period (Figures 1-3). Out of the 13299 included sequences, 4074 (30.6%) clustered with at least one other sequence at the time of sampling. At the last observed time point (31 December 2023), 5415 (40.7%) of all sequences were linked to at least one other sequence. We found that although Plog works accounted for 2572 out of the total number of sequences (19.3%), they accounted for 29% of the clustering sequences (Table 2). On the other hand, game related works account for 4782 (36%) of the total number of sequences, but only 25% of clustering sequences, indicating that the propagation frequency of this subgroup may be relatively low (chi-squared < 0.001).

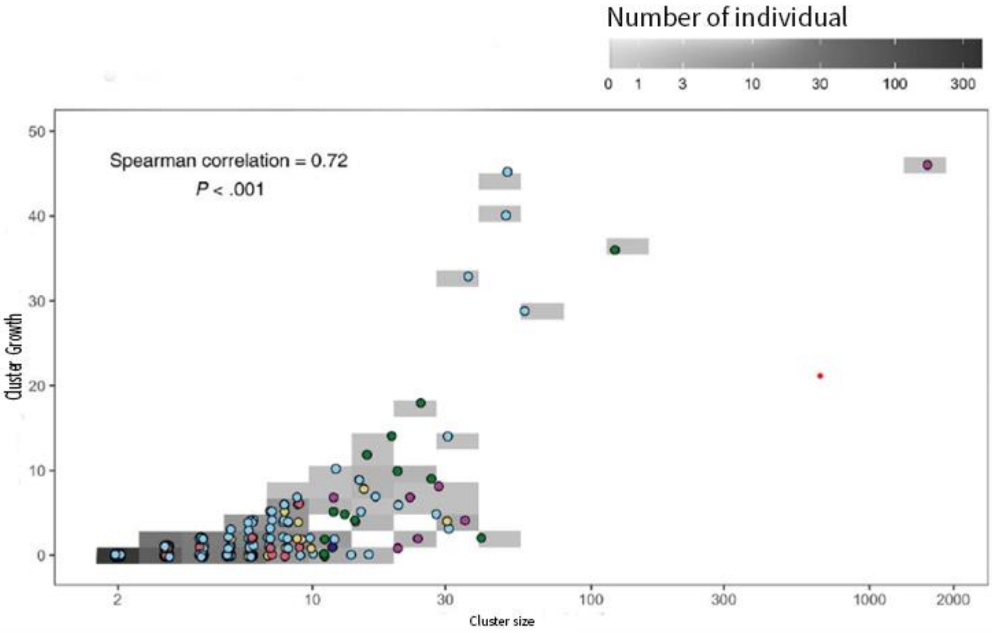

Figure 1: Graph representations of clusters growth and cluster size

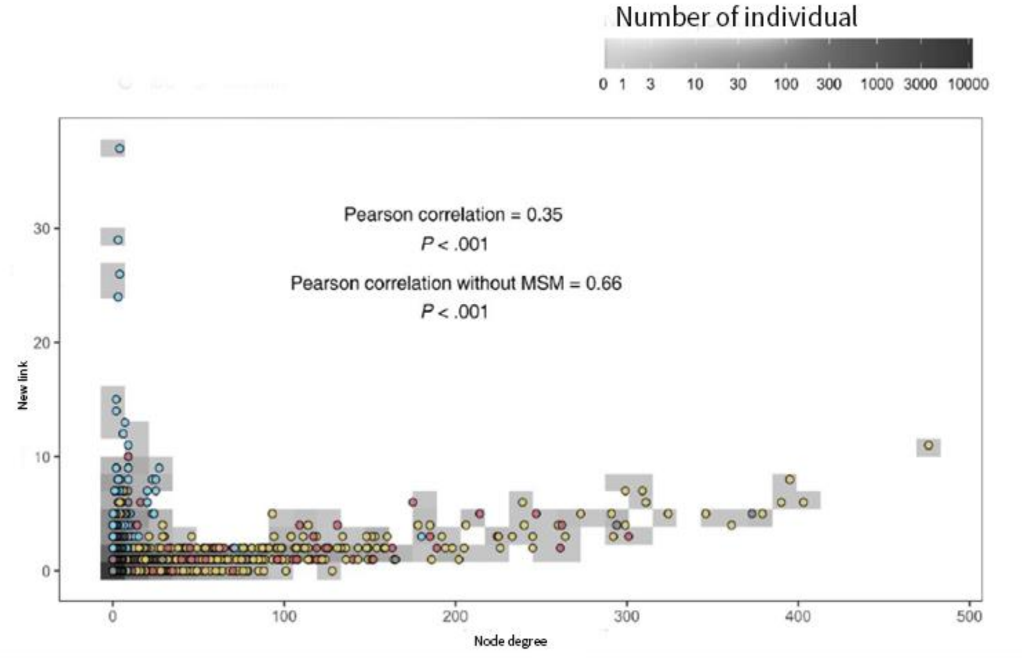

Figure 2: Graph representations of new link and node degree

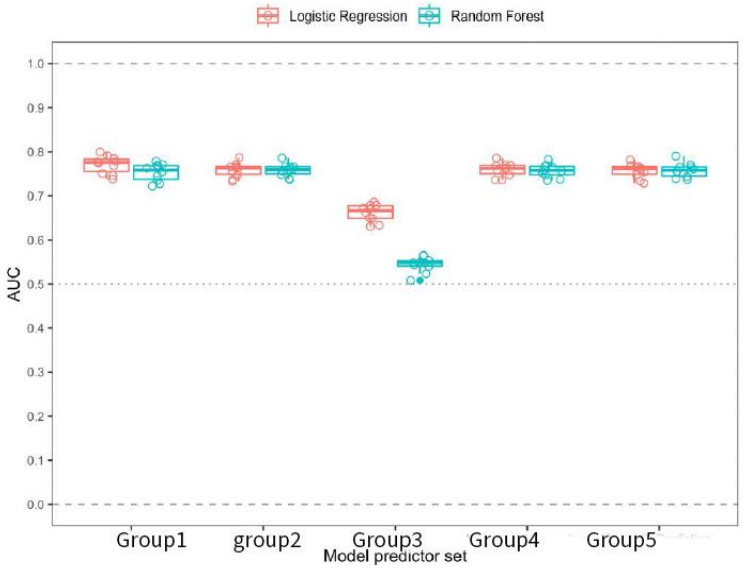

Figure 3: Graph representations of Model Comparison

As of the end of 2023, most clusters had fewer than 10 nodes (943/998, 94.5%). The identified maximum cluster contains 1577 nodes, of which 43.7% are classified as video producers and 24.0% are classified as game content producers [10].

The obtained clusters exhibited a large heterogeneity in terms of composition, size, and growth patterns. Of 575 clusters identified up to 31 January 2023, only 134 (23.3%) gained any new nodes in the following 12 month, of which only 33 (5.7%) gained 5 or more new nodes. Despite the small fraction of clusters that gained 5 or more nodes, they accounted for 443 (70.9%) of all 625 nodes that were gained by all 575 clusters collectively. The clusters that gained 5 or more nodes were disproportionately Beauty content clusters (27 of 33, 81.8%). We found a strong correlation between cluster size in 2023 and number of new nodes acquired up to December 2023 (Spearman r=0.72, P<.001). Similarly, of the 9308 individuals sampled as of January 2023, only 1079 (11.6%) gained links to new sequences up to the December 2023. Most individuals that acquired new links only gained very few: only 206 (2.2%) gained links to 3 or more new sequences, and they accounted for 1103 (50.4%) of the total 2190 new links over the studied period [11].

When modeling cluster growth using Poisson regression (with log10 cluster size in the June 2023 as an offset), we found that past growth of a cluster was a good predictor for future growth (High likes rate ratio [HIRR], 5.11 [95% confidence interval, 95% CI, 2.62–9.95] and HIRR, 11.03 [95% CI, 6.44–18.88] for past cluster growth of 2–3 and ≥4, respectively). Besides past growth, no other variable yielded a statistically significant estimate in the multivariable model. On the other hand, clusters with "Plog or not" had significantly higher growth rates in the univariable model (HIRR, 2.18 [95% CI, 1.41–3.37]). Figure 4 indicates that the effect of these variables can be captured by including past cluster growth as a proxy for predict factors. Similar results were obtained when restricting the analysis to clusters where “Beauty content” was the most common acquisition category and when varying the time period considered [12].

Table 2: Characteristics of the clustering dataset

Characteristic	Clustered $(n = 4074)$	Not Clustered $(n = 9225)$	All $(n = 13299)$
Age			
Median	36(31,42)	37(31,45)	37(31,44)
Sex, No. (%)			
Female	899(22.1)	2844(30.8)	3743(28.1)
Acquisition group, No. (%)			
Beauty content	1751(43.0)	3588(38.9)	5339(40.1)
Games related	1017(25.0)	3765(40.8)	4782(36.0)
Plog or not	1180(29.0)	1392(15.1)	2572(19.3)
Unknown	126(3.1)	480(5.2)	606(4.6)
Comments			
Median	27162(3345,110023)	15900(790,87078)	19605(1260,95938
Favourite count			
Median	379(222,568)	340(180,527)	350(191,540)
Missing, No. (%)	68(1.7)	184(2.0)	252(1.9)

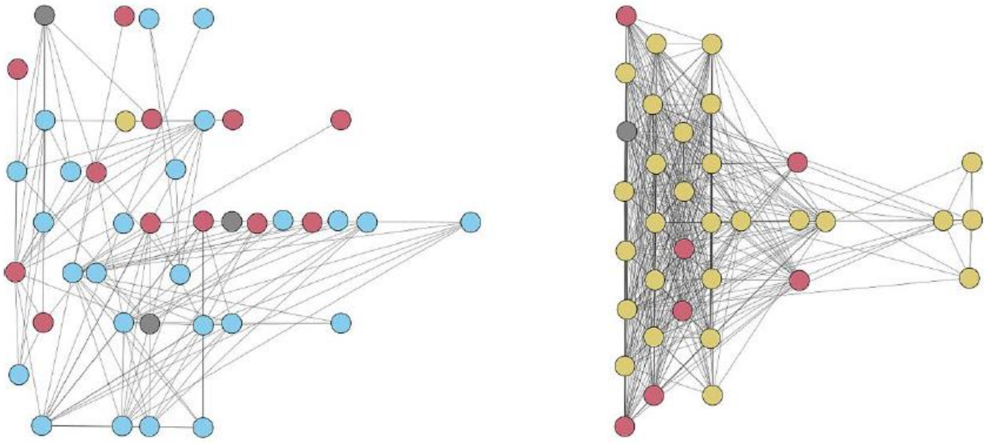

Figure 4: Graph representations of 2 different clusters Red indicates Games related; Yellow indicates Plog or not; Blue represents Beauty content; Gray indicates Unknown

To quantify the relevant factors of growth at the individual node level, that is, a node's change of acquiring new links over time, we specified a logistic regression model where we used a similar set of variables for predicting the addition of new links to a given node within 3 months of being sequenced (Figure 3). Node degree had a significant effect on the outcome, with larger node degrees being associated with higher probabilities to gain new links (odds ratio [OR], 2.41 [95% CI, 1.94–3.00], OR, 4.98 [95% CI, 3.98–6.24], and OR, 11.35 [95% CI, 8.34–15.45] for node degree 1, 2–4, and ≥5, respectively). Accordingly, removing node degree from the regression model led to a significantly worse model fit (likelihood ratio test, P<.001). In other words, the growth of the network occurs by preferential attachment, meaning more connected nodes acquire a new links, which also explains the approximately scale-free pattern observed for the degree distribution of the whole network.

3.7. Model Comparison

We trained multiple models in Table 3, using 5 different sets of predictors with the goal of identifying the best model for predicting whether a certain node is going to acquire a link to a new node within 3 months. To assess the performance of these models, we performed a 10-fold cross-validation and compared the median areas under the curve (AUCs) of the receiver operating characteristic (ROC)-curves based on the model predictions.

Table 3: Regression table

Variable	Attribute	Odds Ratio (95%-Cl)

Node degree (ref=0)	1	3.33(2.94-3.76)
		2.41(1.94-3)

	2-4	8.05(7.06-9.18)
		4.98(3.98-6.24)

	≥5	30.17(25.48-35.73)
		11.35(8.34-15.45)

Age (ref =<40 years)	40-49	0.63(.57-.7)
		0.52(.42-.64)

	≥50	0.48(.42-.55)
		0.82(.56-1.19)

Acquisition group (ref = Beauty)	Evaluation blogger	0.51(.46-.57)
	Tutorial bloggers	0.88(.72-1.07)
Favourite count (ref = <50)	50-999	0.89(.7-1.11)
	1000-9999	1.42(1.17-1.71)
	≥10000	0.95(.69-1.3)

Comment - count (ref =<300)	≥300	1.54(1.42-1.69)
		1.59(1.31-1.93)

Games related (ref = No)	Yes	1.48(1.28-1.72)
		1.13(.94-1.37)

Plog or not (ref = No)	Yes	1.73(1.47-2.02)
		1(.83-1.21)

Among models with preselected predictor sets, models that used both network and patient characteristics yielded the most accurate predictions. Random forests and logistic regression models performed similarly in all cases except one. Notably, restricting the set of predictors to author identity information and Beauty content variables resulted in a large drop in accuracy: from the mix to the individual predictor set, the median AUC decreased from 0.78 to 0.67 for the logistic regression and from 0.76 to 0.55 for the random forest. On the other hand, restricting the set of predictors to variables pertaining to the topological characteristics of clusters and nodes (cluster predictor set) did not decrease accuracy to the same degree, as the median AUC was 0.76 both for the logistic regression and the random forest. Accordingly, variables with the highest variable importance in the mix random forest model were cluster characteristics, namely node degree, past cluster growth, and cluster size.

4. Conclusion & Discussion

We apply a clustering approach based on evolutionary distance to a longitudinal data set of Redbook. We analyze the cluster growth dynamics using statistical learning methods on the Red book dataset .We find that, over the past year's time span, there has been a small increase in the number of Reddit image content creators. Similarly, many creators have not formed any new connections by 2024.Consistent with this, we find that favorites, creator type can predict cluster growth. This suggests that some of the information provided by the above variables can be captured by the network's characteristics. And using the poisson regression model is better than the random forest in the multi-class test. The disadvantage of this analysis is that the number of variables used for prediction is still not enough. In the next step, we will model on a larger dataset.

References

[1] Hung, C.C., Huang, Y.C., Hsu, J.Y.J. and Wu, D.K.C., 2008, July. Tag-based user profiling for social media recommendation. In Workshop on Intelligent Techniques for Web Personalization & Recommender Systems at AAAI (Vol. 8, pp. 49-55).

[2] Lee, C.H., 2012. Unsupervised and supervised learning to evaluate event relatedness based on content mining from social-media streams. Expert Systems with Applications, 39(18), pp. 13338-13356.

[3] Tang, J. and Liu, H., 2014. An unsupervised feature selection framework for social media data. IEEE Transactions on Knowledge and Data Engineering, 26(12), pp. 2914-2927.

[4] Y. Yang, H. Shen, Z. Ma, Z. Huang, and X. Zhou. L21-norm regularized discriminative feature selection for unsupervised learning. In IJCAI, 1589-1594, 2011.

[5] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. NIPS,507–514,2006.

[6] Belkacem, S., Boussaid, O. and Boukhalfa, K., 2020. Ranking news feed updates on social media: A comparative study of supervised models. In EGC (pp. 499-506).

[7] Natekin, A. and Knoll, A., 2013. Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, p.21.

[8] Biau, G. and Scornet, E., 2016. A random forest guided tour. Test, 25, pp.197-227.

[9] Rahman, H., Tariq, J., Masood, M.A., Subahi, A.F., Khalaf, O.I. and Alotaibi, Y., 2023. Multi-tier sentiment analysis of social media text using supervised machine learning. Comput. Mater. Contin, 74, pp.5527-5543.

[10] Rish, I., 2001, August. An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

[11] Song, Y.Y. and Ying, L.U., 2015. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), p.130.

[12] Joachims, T., 1998. Making large-scale SVM learning practical (No. 1998, 28). Technical report.