WEP Warwick Evans Publishing

International Journal of Social Sciences and Public Administration

ISSN: 3005-9836 (Print), ISSN: 3005-9585 (Online) | Volume 5, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijsspa.v5n1.04
Journal homepage: https://wepub.org/index.php/IJSSPA/index

A Big Data-based Approach to Web User Behavior Analysis

-- An Overview of Cluster Analysis

Hongwei Huang

Shandong University, Weihai, 264299, China

ABSTRACT

In China, the Internet has developed to a relatively mature scale, and Internet applications have gradually transitioned from being singular to diversified. The Internet is changing people's ways of learning, working, and living, and even influencing the progress of the entire society. Against the backdrop of rapid Internet development, we have gradually entered the "big data era." Faced with such a vast amount of data, single-machine statistics have become inadequate. This paper first introduces the background and significance of the research, providing a preliminary description of network traffic from the perspective of network services. It then elaborates on the concepts and classifications of network user behavior, along with key data, mainly introducing analysis methods. The primary method used in this paper is cluster analysis, including distance and similarity coefficients in cluster analysis, with a focus on the main steps and algorithm process of k-means clustering. Finally, the paper conducts user behavior analysis based on the clustering results.

KEYWORDS

China Internet; Big Data Era; Network Traffic; Network User Behavior; Cluster Analysis; Distance and Similarity Coefficients; K-Means Clustering; User Behavior Analysis; User Behavior Analysis Clustering Analysis.

1. INTRODUCTION

1.1. Background and Significance of the Study

In recent years, with the rapid advancement of the Internet and cloud computing technologies, internet applications have evolved from simplicity to diversification. Nowadays, people can access a vast array of useful knowledge from the Internet's expansive knowledge base. The swift growth of the Internet has led to the continuous generation of data across various industries.

According to the "China Internet Development Report" released by the China Internet Network Information Center, as of June 30, 2018, the number of Internet users in China had reached 802 million, with 788 million being mobile phone users, accounting for 98.3% of the total. The Internet penetration rate is 57.7%. This clearly demonstrates that the Internet is intricately linked to people's daily lives.

1.2. Research Status Domestically and Internationally

Today, the storage capacity of computers has steadily increased, and algorithms have become increasingly sophisticated. Recent years have seen exponential growth in data. It is estimated that by 2020, the volume of data will reach 40 ZB. Big data holds immense value.

Data mining involves extracting valuable information from large datasets, revealing connections and trends among the data. Effective data mining requires the collection of comprehensive data, as richer data sets yield more accurate insights. By selecting appropriate methods, one can swiftly and efficiently acquire an understanding of users' internet habits and preferences.

1.3. Main Research Content of the Thesis

This thesis focuses on the analysis of network user behavior in the context of big data. The main research content includes the concept and classification of network user behavior, key data points, calculation of similarity coefficients and distances in cluster analysis, k-means clustering, and a detailed understanding of the main steps and algorithmic processes of k-means clustering.

1.4. Organization of the Thesis

The specific chapters of this thesis are arranged as follows:

Chapter 1 mainly introduces the research background, research significance, and current research status of this topic, and finally introduces the structure of this paper.

Chapter 2 introduces the concepts and classifications of network user behavior and key data, mainly introducing analysis methods and analytical approaches, and finally provides a summary.

Chapter 3 introduces cluster analysis. It starts with the concept, using the understanding of similarity coefficients and distances to judge the relationship between samples. It focuses on the main steps and algorithm flow of k-means clustering.

2. BIG DATA AND USER BEHAVIOR ANALYSIS

2.1. The Concept and Classification of Network User Behavior

User behavior involves collecting, integrating, and analyzing all data generated from users using a product (such as visit volume, visit rate, visit frequency, retention time, etc.) to understand the patterns of product usage. This analysis provides robust data support for the next steps in product optimization and marketing.Network users can be broadly categorized into three types[4]:

(1) Classification by User Group Characteristics: Classifying user groups is not limited to single characteristics like age or gender. It requires categorizing from different dimensions after obtaining user behavior data.

(2) Classification by User Product Usage Rate: Network products reflect this in data such as traffic, visit rate, click volume, click rate, visit frequency, and retention time. For mobile applications, this is reflected in download volume, usage frequency, and usage modules.

(3) Classification by User Product Usage Time: Understanding when users use the product throughout the day.

2.2. Key Data of Network User Behavior

This paper will analyze the key data of user behavior, as follows:

(1) Visitor Traffic: This includes user groups, visitors, visit volume, traffic sources, traffic pages, visitor analysis, visitor visit analysis, search engines, advertising effectiveness analysis, malicious click analysis, etc.

a.**User Groups:** The visit volume within 24 hours in the user's region.

b.**Visitors:** The primary sources of visits, such as country, province, city, etc.

c.**Visit Volume:** Analyzing the visit volume at various stages throughout the year to determine the peak periods of the website.

d.**Page Views:** The content viewed by visitors within a specific time period.

e.**Traffic Sources:** Analyzing the sources of the website's traffic.

f.**Traffic Pages:** The main pages that attract the traffic.

g.**Visitor Analysis:** The number of return visits by users within 24 hours, the number of pages viewed by users, and the time spent on the website.

h.**Malicious Click Analysis:** Loss analysis, defense strategies, etc.

(2) Visit Volume by Time Period: Analyzing visit peaks, including when during the day the visits occur and the specific number of visits. By analyzing the common characteristics of time periods, it is possible to determine which user groups are the key consumers during specific time periods.

(3) Analysis of Visit Volume by Time Period for Consumption Content: First, classify the consumption content, then analyze the time periods to identify the consumption times of different user groups, preparing for the next step of targeted marketing.

(4) User Analysis: Includes data on users' age, gender, education, occupation, etc.

2.3. Analysis Methods of Network User Behavior

After collecting big data, user analysis is carried out. The analysis methods are mainly divided into three types:[5].

(1) Data Analysis-Oriented: Data analysis-oriented means that during the process of big data analysis, user recommendation services are designed based on data collection. For example, on e-commerce websites like Taobao and JD.com, recommendations can be made based on predicting what you might like and offering related suggestions.

(2) Product Design Feedback-Oriented: In the analysis process, data collection, organization, and mining are conducted to provide data analysis for product design and improvement. This mainly involves analyzing user attributes and habits.

(3) User Survey-Oriented: Understand users' preferences for the product by analyzing the frequency of their usage.

2.4. Methods of Analyzing Network User Behavior

Due to individual differences, each user has their own habits when surfing the internet. Businesses can conduct targeted marketing by analyzing each user's online behavior, making the correct method of analyzing network user behavior particularly important. Currently, the most popular data analysis methods include cluster analysis, decision trees, hypothesis testing, reliability analysis, etc. The following briefly lists and describes several commonly used data analysis methods.

(1) Descriptive Statistics: Descriptive statistics describe the central tendency, dispersion, skewness, and kurtosis of the data.

(2) Hypothesis Testing: Parametric tests are used to test major parameters (such as mean, percentage, variance, correlation coefficient, etc.) under the assumption that the population distribution is known.

(3) Reliability Analysis: Reliability analysis refers to checking the credibility of measurements, such as the authenticity of survey questionnaires.

(4) Contingency Table Analysis: Contingency table analysis can be used to determine whether there is a correlation between variables.

(5) Correlation Analysis: Primarily investigates whether there is a dependency relationship between phenomena and explores the direction and degree of correlation for those phenomena that have a dependency relationship.

(6) Analysis of Variance (ANOVA): Conducting ANOVA requires certain conditions: each sample must be an independent random sample; each sample must come from a normally distributed population; and the variances of the populations must be equal.

(7) Cluster Analysis: Cluster analysis refers to methods of classifying individual samples or index variables based on their characteristics, finding reasonable statistics to measure the similarity of entities.

(8) Other Analysis Methods: Other analysis methods include multiple response analysis, distance analysis, item analysis, correspondence analysis, decision tree analysis, neural networks, system equations, Monte Carlo simulation, etc.

2.5. Summary

This chapter first provides a simple classification of network users, then interprets the data generated by network user behavior. Next, it conducts a preliminary study on the methods of analyzing network user behavior, and finally, introduces some current methods of user behavior analysis. In Chapter 3 of this paper, the cluster analysis method will be introduced in detail.

3. CLUSTER ANALYSIS

Cluster analysis refers to the method of classifying individual samples or index variables based on their characteristics, and finding reasonable statistics to measure the similarity of entities.

3.1. Overview

Cluster analysis is the process of grouping samples that do not have inherent categories into different clusters and describing each cluster.[6]As shown in Figure 1, the data can be divided into three different clusters: red, blue, and green, each with its unique characteristics.

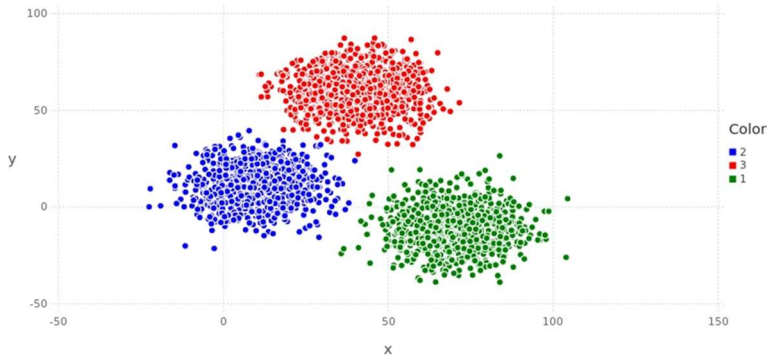

Figure 1. Data Classification

3.2. Measurement of the Proximity between Samples or Variables

To study the proximity between samples or variables, there are two quantitative standards:

(1) Similarity Coefficient: The similarity coefficient measures the degree of similarity between attributes or samples. A similarity coefficient closer to 1 or -1 indicates a higher similarity between

the variables or samples. Conversely, a similarity coefficient closer to 0 indicates more independence between the variables or samples. Similar samples are classified into the same cluster, while different samples are classified into different clusters.

(2) Distance: Each sample is considered a point in a P-dimensional space, and some metric is used to measure the distance between points. Points that are closer together are more likely to belong to the same category, while points that are further apart are likely to belong to different categories.

3.3. Cluster Statistics for Ratio Variables

3.3.1. Cluster Statistics for Ratio Variables - Distance Metrics

(1) Manhattan Distance: Also known as city block distance, it is the sum of the absolute differences of the coordinates of two points on a standard coordinate system.

$$d_{ij}(1)=\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right| \tag{1}$$

(2) Euclidean Distance: It is a commonly used distance metric. The calculation formula is:

$$d_{ij}(2)=\left[\sum_{k=1}^{p}(x_{ik}-x_{jk})\right]^{\frac{1}{2}} \tag{2}$$

(3) Minkowski distance: The Minkowski distance is a generalization of the Manhattan and Euclidean distance formulas. Its calculation formula is:

$$d_{ij}=\left[\sum_{k=1}^{p}(x_{ik}-x_{jk})^{q}\right]^{\frac{1}{q}} \tag{3}$$

(4)Canberra Distance: The Canberra distance is a dimensionless measure. Its calculation formula is:

$$d_{ij}(L)=\sum_{k=1}^{p}\frac{\left|x_{ik}-x_{jk}\right|}{x_{ik}+x_{jk}} \tag{4}$$

(5)Mahalanobis Distance: Unlike the Euclidean distance, the Mahalanobis distance takes into account the relationships between different characteristics. Its calculation formula is:

$$d_{ij}(M)=\left[(x_i-x_j)'S^{-1}(x_i-x_j)\right]^{\frac{1}{2}} \tag{5}$$

3.3.2. Cluster Statistics for Ratio Variables - Similarity Coefficient Metrics

(1) **Correlation Coefficient:** The correlation coefficient reflects the degree of correlation between two variables by multiplying their deviations. Its calculation formula is:

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x_i})(x_{kj} - \overline{x_j})}{n} \tag{6}$$

(2) Cosine Similarity: When the angle between two vectors is 0°, the cosine similarity is 1, indicating that they are extremely similar. When the angle is 90°, the cosine similarity is 0, indicating that they are not related. Its calculation formula is:

$$C_{ij} = \frac{\sum_{k=1}^{n} x_{ki} x_{kj}}{\left[\left(\sum_{k=1}^{n} x_{ki}^{2}\right)\left(\sum_{k=1}^{n} x_{kj}^{2}\right)\right]} \tag{7}$$

3.4. Types of Cluster Analysis

Based on different clustering methods, clustering can be divided into hierarchical clustering and K-means clustering.

(1) Hierarchical Clustering refers to a clustering process that follows a certain hierarchy. It builds a tree-like structure by either merging smaller clusters into larger ones (agglomerative) or splitting larger clusters into smaller ones (divisive) based on the similarity of the data points.

(2) K-Means Clustering Algorithm randomly selects K objects as the initial cluster centers, then calculates the distance between each object and each of the seed cluster centers. Each object is assigned to the cluster center closest to it. The cluster centers and the objects assigned to them represent the clusters. After all objects are assigned, the cluster centers for each cluster are recalculated based on the current objects in the cluster. This process continues to repeat until the termination condition is met. The termination conditions can be:

1)No (or a minimal number of) objects are reassigned to different clusters.

2)No (or a minimal number of) cluster centers are recomputed.

3)The sum of squared errors reaches a local minimum.

This iterative process ensures that the clusters formed are as compact and well-separated as possible based on the specified criteria.

3.4.1. Hierarchical Clustering

(1) Common Methods for Inter-Cluster Distance Measurement

Common methods include Nearest Neighbor, Furthest Neighbor, Between-Group Linkage, Within-Group Linkage, Centroid Clustering, Median Clustering, and Sum of Squares of Deviations.

Nearest Neighbor Method. As shown in Figure 2, the distance between the two closest individuals of two different types is taken as the inter-cluster distance.

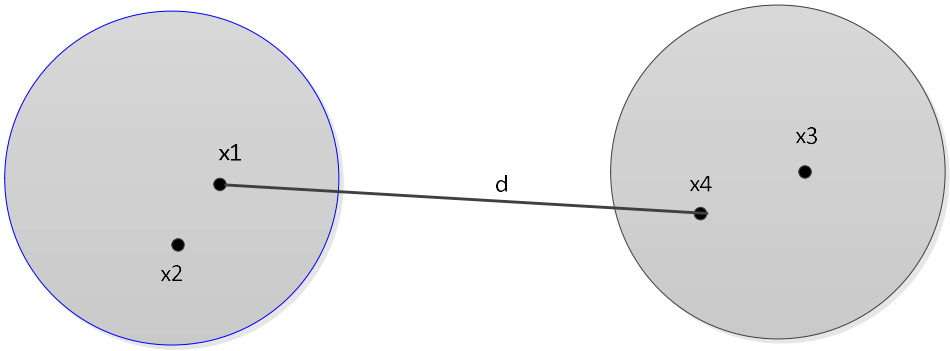

Figure 2. Nearest Neighbor Method

Furthest Neighbor Method. As shown in Figure 3, the distance between the two furthest individuals of two different types is taken as the inter-cluster distance.

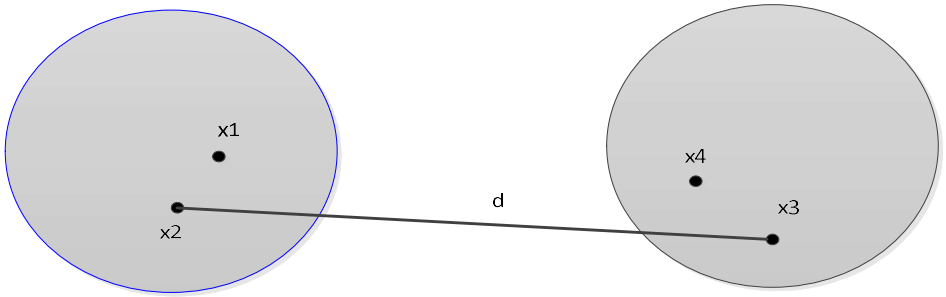

Figure 3. Furthest Neighbor Method

Between-Group Linkage Method. As shown in Figure 4, the average of the distances between all pairs of individuals from the two clusters is taken as the inter-cluster distance.

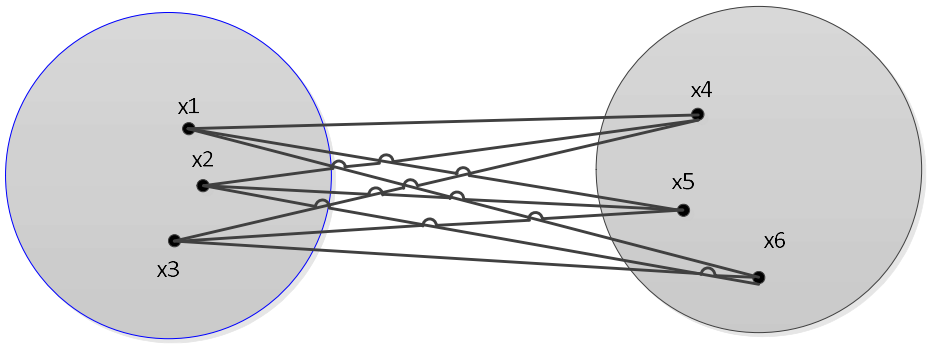

Figure 4. Between-Group Linkage Method

$$d = \frac{d_1 + d_2 + d_3 + d_4 + d_5 + d_6 + d_7 + d_8 + d_9}{9} \tag{8}$$

Within-Group Linkage Method. As shown in Figure 5, the two clusters are merged into one, and the average distance between all individuals in the combined cluster is taken as the inter-cluster distance.

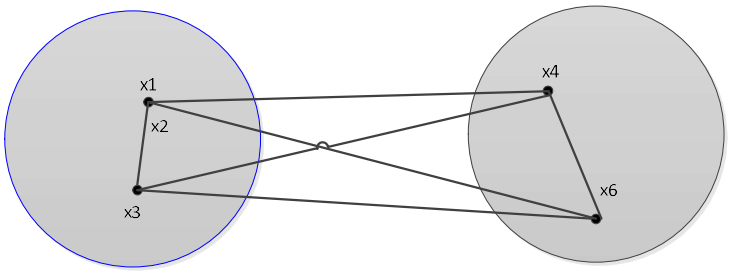

Figure 5. Within-Group Linkage Method

$$d = \frac{d_1 + d_2 + d_3 + d_4 + d_5 + d_6}{6} \tag{9}$$

Median Clustering Method. The distance between the medians of the two types of variables is used as the inter-cluster distance.

(2) Main Steps of Hierarchical Clustering

Hierarchical clustering mainly consists of five specific steps: selecting variables, data transformation and preprocessing, cluster statistical computation, clustering, and interpretation and validation of clustering results.

3.4.2. K-Means Clustering

K-means clustering requires the number of clusters to be specified in advance. Its main characteristic is that it performs particularly fast for large sample sizes.

(1) Main Steps of K-Means Clustering

1) Create an initial partition by randomly selecting k elements from the set U to serve as the centers of the k clusters.

2) Calculate the similarity of the remaining elements to each of the k cluster centers and assign each element to the cluster with the highest similarity.

3) Based on the clustering results from the previous round, recalculate the center of each of the k clusters. This is done by taking the arithmetic mean of each dimension of all elements within the cluster.

4) For all elements in U, excluding the new cluster centers, re-cluster according to the method in step 2).

5) Repeat steps 3) and 4) until the difference between the current clustering result and the previous clustering result is less than a set threshold.

6) Output the final clustering results.

REFERENCES

[1] 42nd Statistical Report on the Development of the Internet in China, July 2018 by the China Internet Network Information Center (CNNIC).http://www.cnnic.cn/.

[2] Ren, S. Y. (2014). Analysis of Internet User Behavior Based on Big Data [Master's thesis, Beijing University of Posts and Telecommunications]. Pages 18-20. December 2014.

[3] Hu Yanqing, Zhou Jinyan, Xu Xiaona. Application of Data Mining in Mobile User Behavior Analysis System JJJ. Modern Telecommunication Technology. 2013.01, 01(01): 86-89.

[4] Zuo Jun. Network User Behavior Analysis Based on Big Data JJJ. Software Engineer, 2014.05, 39(8): 556-558.

[5] Chen Wenwei. Overview of Data Mining and Knowledge Discovery JJJ. Computer World News, 1997.05, 24(8): 122-124.

[6] Dong Fuqiang. Research and Application of Network User Behavior Analysis DDD. Master's Thesis, Xi'an: Xidian University, 2005.01, pp. 11-12.

[7] Gordon S. Linoff, Michael J. A. Berry. Mining the Web: Transforming Customer Data into Customer Value. Publishing House of Electronics Industry, 2004, p. 3.

[8] Wang Shi, Gao Wen, Li Jintao, et al. Knowledge Discovery of Path Clustering in Web Sites JJJ. Journal of Computer Research and Development, 2001.04, 4(4): 482-485.

[9] Niu Wenjia, Liu Jiqiang, Shi Chuan, et al. User Network Behavior Profiling MMM. Beijing: Publishing House of Electronics Industry, 2018.05, pp. 30-36.