VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY
UNIVERSITY OF TECHNOLOGY
FACULTY OF COMPUTER SCIENCE AND ENGINEERING

BIG DATA CLUB PROJECT PROPOSAL

Data processing, analysis, visualization and mining user data via Content Delivery Network system

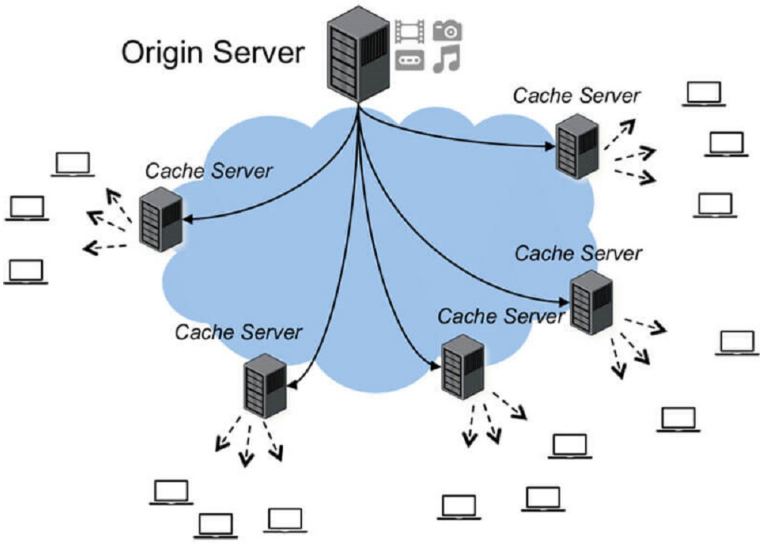
Team:	7
Students:	Lê Bảo Khánh (Leader)
	Trần Ngọc Mai Thảo
	Kha Sang
	Nguyễn Đình Thiên Huy
Email:	khanh.lehcmut@hcmut.edu.vn

HO CHI MINH CITY, MARCH 2022

Contents

1 Introduction

With the development of the Internet, the demand for online recreation has increased over time. Many platforms have been developed creating a vibrant yet competitive industry. In order to survive, companies have to upgrade their products to amaze their clients such as releasing new models, adding unique features or even advertising. However, what really matters is the quality of the platform itself and the customers' experience, especially when it comes to online contents which account for a large amount of Internet traffic but require high load speed such as video streaming or online games. In order to optimize load speed, service providers have to apply many technical methods, and one of the most commonly-used solutions is setting up a geographically distributed group of servers which work together, which is usually referred to as Content Delivery Network (or CDN in short).



Not only does CDN tackle with network congestion but also bring in a massive amount of data created by users through the platforms. From the contents aspect, being acknowledged about the popularity of each product will support companies in adopting appropriate strategies and adjusting the recommendation system to suit with each customer. For internet service, such information will help them to distribute the servers properly to satisfy the customers and optimize the budget for this job.

In this paper, we will discuss further about the benefits of extracting and analyzing user data from CDN, our method of approaching this issue as well as a few relevant researches.

2 Motivation

2.1 Research aim

Internet service quality perceived by customers is largely unpredictable and unsatisfactory. Content Distribution Network (CDN) is an effective approach to improve Internet service quality. This research is firstly conducted to understand about the CDN and how it works. We began a search on the definition of CDN, CDN's operation, and assessed the way to analyse raw data and predict the content popularity. In this way, our team can understand the situation and know the reasons why CDN is used for almost websites and social media, as well as the process to make prediction for content popularity, ranging from data pre-processing, analysis, visualization, and mining.

2.2 Research objective

When doing research, we start with the rationale for understanding how CDN comes about. It is also necessary to consider the internet situation of Vietnam. We then give an overview about CDN, show some outstanding research about CDN, our expected product, and the approaches proposed in literature to conduct this research. An example of CDN is described to show how a real commercial CDN operates. We conclude with a brief projection about CDN.

2.3 Expected product

After being processed, data will be visualized with graphs for easy understanding by viewers. Then content popularity is then predicted by linear regression. This presentation can help managers take a deeper look in their business process on online marketing, and make more precise strategies in the future for the sake of their company's development.

3 Relevant related work

Due to the unstoppable growth of data from users logging into webpages, getting and extracting the information from those websites has been an important step for any data analyst if they want to know "the value" of that website, or the whole company at large. Teevan et al.[1] showed, via query log analysis, that nearly 40% of queries were attempts to re-find previously encountered results. In fact, there are numerous researches into web log data mining, ranging from basic data analytics techniques to more complicated approaches (e.g. Machine Learning). Therefore, it is essential that we should dive into this topic, in order to visualize historical data and discover website patterns created by users, which were written on the server log files. Also, we will further study a website's security using several log mining techniques.

According to Qingyu Zhang et al.[2], web mining include three subsections: Web content mining, Web usage mining, Web structure mining; either any of three areas has had ground-breaking contributions to the Data Mining Society. However, in the context of log analysis, we will only focus on web usage mining most popular researches.

Mobasher et al.[3] were the pioneers of automatic web personalization, created a software called WebPersonalizer based on web usage mining, which include end-to-end procedure: data preparation, usage mining, online recommendation. In the first stage, the raw data collected was cleaned with filtering support. To the next step, association rule and clustering algorithms were utilized to discover users favorite item. After all, the model, a.k.a personalized recommendation is proposed.

Srivastava et al.[4] instantiated web usage mining as the application of data mining to find hidden patterns, which include three phases:

- Preprocessing: data cleaning, user and session identification.
- Pattern Discovery: statistical analysis, association rules, cluster or classification.
- Pattern Analysis: the patterns are evaluated by human knowledge.

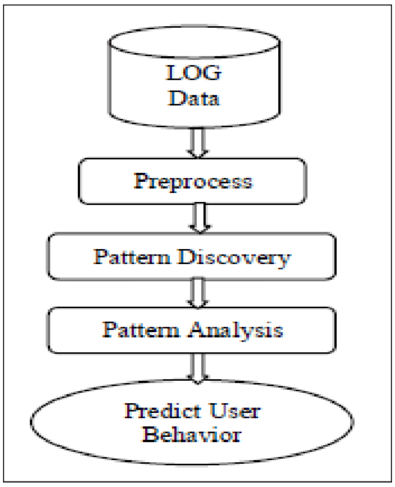


Figure 3.0.1: Web Log Mining system structure

Song and Shepperd [5] discuss the importance of web user clustering path recognition as a directed graph using vector analysis and fuzzy set theory to cluster users and URLs. The effect of this technique is crucial in analysing E-Commerce.

Those are the fundamental studies into the foundation of web usage mining. Recently, Arvind and Gupta [7] propose a methodology to classify users navigation patterns and also to predict the behaviour and interest of the website users. In particular, the reports undergo a time analysis and pageview analysis.

An overview of soft computing frameworks such as neural network, genetic algorithms used in web mining described by S.Pal et al.[9]. The limitations of some of the existing Web mining methods and tools are enunciated, and the significance of soft computing are highlighted. In another approach, Sandro et al.[10] shows how a combined use of data from data warehouse and web data can contribute to improve marketing activities.

M.T. Nguyen et al.[11] propose a novel model to construct and add new attributes encompassing country, province (or city), Internet Service Provider (ISP) from the existing attribute IP given from the log file. . Such knowledge can be leveraged for optimizing system performance as well as enhancing personalization. Furthermore, the valuable knowledge can be useful for deciding reasonable caching policies for web proxies.

Recent years have seen the popularity of using Content Delivery Networks (CDNs) to handle network congestion. H.L. La et al.[12] use real data log file from a large CDN solution vendor in Vietnam in order to take a closer look at the design,implementation, solution, and performance of a Content Delivery Network system by analyzing its raw log file, therefore o understand user access patterns, the sources of requests, system performance, and how such information can be used to improve the whole CDN system.

4 Methodology

Desk research: apply synthesis methods, analytical techniques to process collected data and statistic (both Vietnamese and foreign language documents) and apply Benchmarking method to compare the analysis's result with research results already available at home and abroad.

Methods of data collection: data used in the study is primary data which is obtained from reputable sources.

In general, our team will process data through the following steps:

- Collection: synthesize and measure information by different collection methods.
- Ingestion: transporting data from one or more sources to a target site for further processing and analysis.
- Preparation: processing and transforming raw data before processing and analysis, including data reformatting and combining datasets.
- Computation: perform arithmetic, statistical, or logical operations on data to describe, and evaluate data.
- Presentation: use graphs, text, tables,... to make it easier for anyone to visualize the results, even those who are not data experts. From there, decision making will also be more informed.

5 Plan

We are going to implement our project in 2 months with 5 main phases:

Index	Phase	Timeline	Goals
1	Project Proposal	16/2 - 15/3	Introduce our project, purpose. related work. methodology, plan
2	Data preparation	16/3 - 31/3	Data collection, data ingestion, data preparation
3	Data Exploration	1/4 - 5/4	Gain an insight into the dataset. Determine the research path
4	Data visualization	6/4 - 15/4	Visualize the dataset with graph, table and their relationship
5	Data prediction	16/4 - 30/4	Build a ML model for content popularity prediction

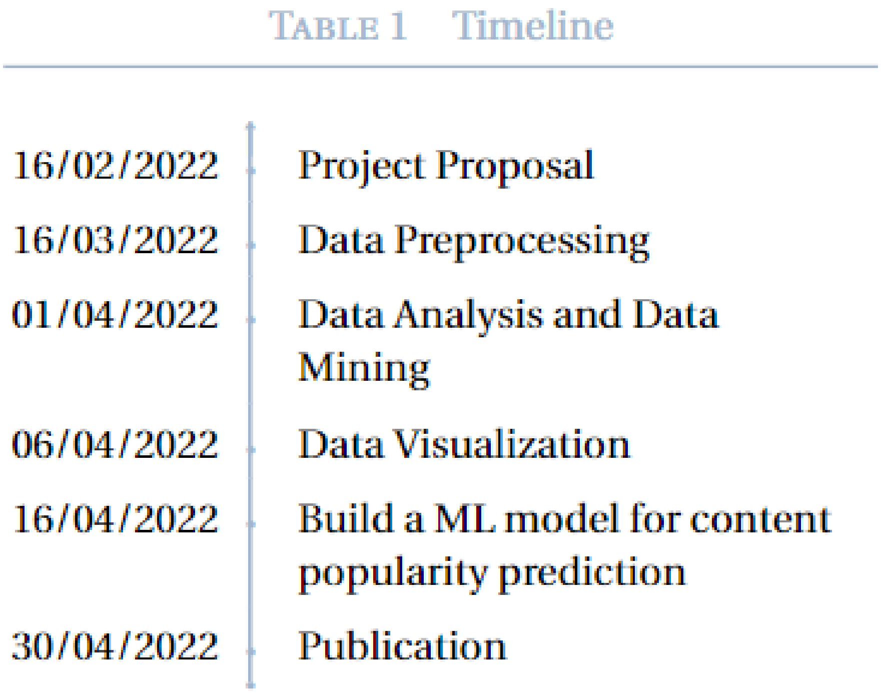


TABLE 1 Timeline

16/02/2022	Project Proposal
16/03/2022	Data Preprocessing
01/04/2022	Data Analysis and Data Mining
06/04/2022	Data Visualization
16/04/2022	Build a ML model for content popularity prediction
30/04/2022	Publication

6 Website

Github:
https://baokhanhle123.github.io/BDC_Assignment/
For more information, please visit our website:
https://baokhanhle123.github.io/BDC_Assignment/

References

[1] Teevan, J., Adar, E., Jones, R., and Potts, M. "Information Retrieval: Repeat Queries in Yahoo's Logs" *SIGIR, 2007*

[2] Qingyu Zhang and Richard S.Segall, "Web Mining: A survey of current research, techniques, and software" *International Journal of Information Technology and Decision Making, December 2008*

[3] B. Mobasher, R. Colley and J. Srivastava, "Automatic personalization based on web usage mining" *Commun. ACM, 2000*

[4] M. Spiliopoulou, "Web usage mining for web site evaluation", *Commun. ACM, 2000*

[5] Q. Song and M. Shepperd, "Mining web browsing patterns for e-commerce" *Comput. Indus, 2006*

[6] Santosh Kumar, Ravi Kumar, "A Study on Different Aspects of Web Mining and Research Issues" *ICCRDA, 2020*

[7] Arvind K.Sharma, P.C.Gupta, "Predicting the Behaviour and Interest of the Website Users through Web Log Analysis" *International Journal of Computer Applications Volume 64– No.7, February 2013*

[8] Navin Kumar Tyagi, A. K. Solanki, Manoj Wadhwa, "Predicting the Behaviour and Interest of the Website Users through Web Log Analysis" *IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 8, July 2010*

[9] S. Pal, V. Talvar, and P. Mitra, ”Web mining in softcomputing framework: relevance, state of the art and future directions” *IEEE Transactions of Neural Networks, 2002, 13(5), pp.1163-1177.*

[10] Sandro Araya, Mariano Silva, Richard Weber, "A methodology for web usage mining and its application to target group identification" *Fuzzy Sets and Systems, Volume 148, Issue 1, 16 November 2004*

[11] Minh-Tri Nguyen, Thanh-Dang Diep, Tran Hoang Vinh, Takuma, Nakajima, and Nam Thoai, "Analyzing and Visualizing Web Server Access Log File" *5th International Conference, FDSE 2018*

[12] Hoang-Loc La, Anh-Tu Ngoc Tran, Quang-Trai Le, Masato Yoshimi, Takuma Nakajima and Nam Thoai, "A use case of Content Delivery Network raw log file analysis" *2020 International Conference on Advanced Computing and Applications (ACOMP) (2020): 71-78.*

[13] Gang Peng, "CDN: Content Distribution Network", Cornell University, *2004*