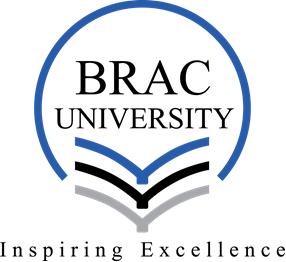
**Pre-thesis -I Report**



Deep learning based predictive analytics for efficient content caching in edge network

Name and ID:

1. Dhruba Chakraborty -18101028
2. Maisha Hossain -18201184
3. Mahima Rabbi - 18101563
4. Maria Khanom Oishi - 17301029
5. Saraf Noor Khaled - 18141009

**Supervisor: Md. Golam Rabiul Alam**

**Co-supervisor:**

**Date of Submission: 10/1/2021**

**Department of Computer Science and Engineering**

**BRAC University**

Deep learning based predictive analytics for efficient content caching in edge network

**Abstract**

Content centric network is a state-of-the-art networking architecture for content distribution and content caching. However, it is inefficient to cache every content in each network devices. The modern edge computing technology opens the door for content caching in the edge of the network. However, still we have to decide which contents we should cache and which content we should replace from the cache. The deep learning based predictive analytics can play an important role in selecting contents for caching purpose. In this research, we will use LSTM based Recurrent Neural Network for predictive content caching at the edge of the network.

1. **Introduction**

Soon after the invention of the first computer ENIAC in 1946, one of the most significant lacking it had was networking. People could do many things with the computer. But, it was impossible to share their works with others who were miles away. From this hunger of sharing, people started to think about making a system by which they could share their works with others. From this consequence, in 1960 ARPANET (The Advanced Research Projects Agency Network) was built in order to create a network with thousands of computers. And, thus the journey of networking had started.

In the very first era of networking, it was just a connection between computers for sharing mostly research data or important files. Only some of the sophisticated researchers and high-level people got to have the benefit of networking. But, in modern time, the concept of networking has changed a lot. Nowadays, there are thousands of fields in networking. People from every stage in the society get help of networking in their day to day life. In this context, content has become the most powerful weapon in the networking field. People use contents to get their job done in their daily life. Starting from media streaming sites, social networking sites, online news portals and many others are spreading digital wellbeing to the human beings through contents.

Content centric network is getting richer day by day with the help of thousands of content providing sites and its users. However, this won’t have been this rich, if it wouldn’t have been efficient. Efficiently caching the contents are so much important in networking. Caching a content means fetching the content from the server. It might be any server all over the world. But, that might be problematic as the server from which the files are being cached, might be far away from the user. That’s where efficient content caching comes in handy. In efficient content caching, files get fetched from the closest server. As a result, lots of time gets saved.

However, there is a significant issue when deciding which content we should cache and which we should replace from the cache because of limited cache memory. We need to cache contents that are more important to the users. But, it is harder to decide which content is more important to the user. To make the purpose easier, we can use deep learning based predictive analytics. Predictive analytics can help us to decide which file to cache and which file to replace from the cache depending on its importance.

1. **Research Problem**

With the mass availability of devices like mobile phones, computers etc., the user of internet is increasing rapidly day by day. And, content providing sites like YouTube, Netflix, Prime Video etc. are becoming so popular among the users. However, people want to stream their contents faster from the sites with less latency. If the requested files are available on the cache server, they are delivered to the users extremely faster. Which is why caching is necessary. Assume a [1] Netflix subscriber in London wants to stream the show House of Cards. To ensure fast access and minimum buffering time, Netflix copies the videos from their origin servers in Los Gatos, CA, to the caching server closest to London. Because of this, all subscribers in London can quickly access the show and avoid a transatlantic file transfer. However, it is impossible to keep every movie in the closest caching server of London because of space limitation. To save the space of the cache server, the not so popular movies are needed to be replaced from the cache server with new ones. As a result, there comes a decision taking between what movies to keep and what movies are needed to be replaced from the cache.

Therefore, a question might arise:

***“Which contents we should cache and which contents we should replace?”***

This research will answer the above question through the usage of Deep Learning based Predictive Analytics Algorithm (in our case, LSTM based RNN).

1. **Research Objectives**

We are going to build a system using deep learning based predictive analytics so that we can decide which contents are needed to be cached and which contents are needed to be replaced from the server. The contents that are trending should be kept in the cache and others should be replaced from the cache. The objectives of the research are:

1. To understand, what content caching is and how it works
2. Important of efficient content caching and its mechanism
3. Importance of edge computing and edge network in efficient content caching
4. To develop a model for connection between predictive analytics and efficient content caching
5. To evaluate the model
6. **Literature Review**

As the blessings of modern technologies like mobile phones, tablets, computers etc. are becoming more affordable and easier to get, people are getting more and more used to with these devices’ day by day. And, people are getting more comfortable with content providing sites like Netflix, YouTube, Prime Video and so on. And, the number of users is rapidly increasing day by day. According to [2] Statista, the number of Netflix users in 2020 is 195.15 million by Q3. However, in a recent article of [3] TNPS (The New Publishing Standard), in 2030 the number of Netflix users is expected to increase up to 500 million.

There might be one problem with the loading time of the contents that are far away from the user. To solve that issue, the concept of caching comes in handy. But, the amount of cache memory is limited. That’s why there is a trade-off between which content to cache and which to replace. To efficiently cache data, predictive analytics is so necessary.

* 1. **Efficient Content Caching**

Content caching is a performance optimization mechanism in which data is delivered from the closest servers for optimal application performance. According to [4] ‘interserver’, when a system accesses the website, the contents in that site will be provided by a nearby cache server rather than the original server which is remote. As a result, it will decrease the latency. However, it is impossible to cache each and every content in the cache server. That’s why efficient content caching is needed. In efficient content caching, most popular contents are cached and least important contents get replaced from the cache server. It reduces server traffic and the performance of the application gets improved.

* 1. **Edge Computing and Edge Network**

Edge networking is a distributed computing paradigm that brings computation and data storage as close to the point of request as possible in order to deliver low latency and save bandwidth.However, [5] edge computing is a modern technology on data center and cloud computing architectures to help create efficiencies. Edge computing is significantly important outside the cloud, at the edge of the network, and more significantly in application where real-time data processing is required. Due to the proximity of the analytical resources to the end users, sophisticated analytical tools and Artificial Intelligence tools can run on the edge of the system. According to [6] ‘The Emergence of Computing’, this placement at the edge helps to increase operational efficiency and contributes many advantages to the system.

* 1. **Predictive Analytics**

Predictive analytics is the use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data. The goal is to go beyond knowing what has happened to providing a best assessment of what will happen in the future [7]. Though predictive analytics has been around for decades, it's a technology whose time has come. More and more organizations are turning to predictive analytics to increase their bottom line and competitive advantage. According to PredictiveAnalyticsToday [8], it uses a number of data mining, predictive modeling and analytical techniques to bring together the management, information technology and modeling business process to make prediction about future. The patterns found in historical and transactional data can be used to identify risk and opportunities for future.

* 1. **Related Works**

This part aims to critically review previous relevant works in the field of Predictive Analytics in the context of Efficient Content Caching at edge network. Observing different techniques used in different relevant research works, we found many challenges in so efficiently caching the contents through prediction.

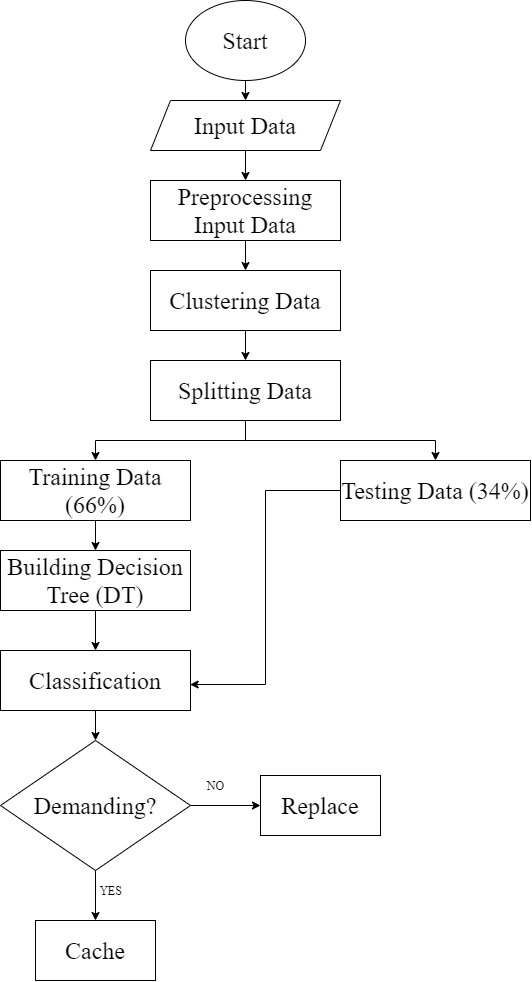
Content caching in edge of the network is so important because if not cached, the data will be accessed by the user directly from the main server through the cloud. Which will increase the latency. According to [9], popular content and objects can be stored and served from edge locations, which are closer to the end users. This operation is also beneficial from the end user perspective since edge caching can dramatically reduce the overall latency to access the content and increase in the sense of overall user experience.

Again, edge computing is another factor in terms of content caching. According to [10], using the cloud as a centralized server simply increases the frequency of communication between user devices, such as smartphones, tablets, wearables and gadgets, referred to as edge devices, and geographically distant cloud data centers. This is limiting for applications that require real-time response. Hence, there has been a need for looking ‘beyond the clouds’ towards the edge of the network, referred to as edge computing. Computing on edge nodes closer to application users could be exploited as a platform for application providers to improve their service. Although, the cache memory at the edge of the network is limited. So, we have to take decision about what content to cache and what content to replace from the cache. That’s where deep learning based predictive analytics comes useful.

Recurrent Neural Network (RNN) is significantly useful for solving the efficient content caching prediction problem because it not only utilizes the current state but also use the previous state data using sequence. According to [9], Unlike the hidden neuron in FNN, the output of RNN depends on both current output of precious layer and last hidden state. However, using RNN might not be appropriate in some cases as there might be data vanishing gradient problem. To solve that issue, LSTM (Long Short-Term Memory) comes handy. According to [11], LSTM models are quite popular due to their special design property related to carefully avoiding vanishing and exploding gradient problem when building deep layer neural network models.

1. **Work Plan**

The aim of using predictive analytics for efficient content caching at the edge network is to cache the most popular contents at the edge of the network and thus decrease the latency. With a view to doing so, the model requires designing a process that takes data from the activity of the users as an input. Then it systematically processes the input data and outputs either of the two different results: ‘cache’ or ‘replace’. The below figure provides a high-level view of the model design:



**Figure 1**: *The flow chart of the predictive analytics model*

We are using LSTM based Recursive Neural Network (RNN) for solving our problem. We could use other Deep Learning based models for this work. But, unlike many other algorithms, LSTM based RNN remembers the previous sequence by keeping them into their memory. As a result, the output gets more and more accurate day by day. The workflow will be done in the following stages:

1. Input data:

In this stage the program takes activity data from the users as input.

1. Input data preprocessing:

In this stage the input data gets formatted in such a way that LSTM can use it to process easily.

1. Processing:

In this stage the formatted input data gets clustered into groups. After clustering, the preprocessed input data is split into two groups; one group is used for training and building the decision tree, and the other group is used for testing the accuracy of the decision tree.

1. Predictions:

In this stage DT is used for prediction to decide whether to cache or replace from cache.

1. **Conclusion**

Predictive analytics is highly important for efficiently caching the contents at the edge of the network. Predictive analytics allows organizations to become proactive, forward-looking, anticipating outcomes and behaviors based upon the data and not on guess or assumptions. So is applicable for content caching. Predictive analytics helps the content providers to cache the most popular contents at the edge of the networks so that users can access them faster with lower latency.

References

1. [What is Content Caching? - Articles for Developers Building High Performance Systems (stackpath.com)](https://blog.stackpath.com/glossary-content-caching/)
2. [• Netflix: number of subscribers worldwide 2020 | Statista](https://www.statista.com/statistics/250934/quarterly-number-of-netflix-streaming-subscribers-worldwide/)
3. [Netflix targets 500 million subscribers by 2030. Rolls out mobile-only subscription option for Nigeria’s 126 million internet users. Publishers look the other way - The New Publishing Standard](https://thenewpublishingstandard.com/2020/10/05/netflix-targets-500-million-subscribers-by-2030-rolls-out-mobile-only-subscription-option-for-nigerias-126-million-internet-users-publishers-look-the-other-way/#:~:text=A%20half%20billion%20subscribers%20by,from%20Digital%20TV%20Research%20shows.)
4. [What is Content Caching? - Interserver Tips](https://www.interserver.net/tips/kb/what-is-content-caching/)
5. [Archive | Design Automation Conference (dac.com)](http://www2.dac.com/events/videoarchive.aspx?confid=170&filter=keynote&id=170-103--0&#video)
6. [The Emergence of Edge Computing - IEEE Journals & Magazine](https://ieeexplore.ieee.org/abstract/document/7807196) [Predictive Analytics: What it is and why it matters | SAS](https://www.sas.com/en_us/insights/analytics/predictive-analytics.html)
7. [What is Predictive Analytics ? (predictiveanalyticstoday.com)](https://www.predictiveanalyticstoday.com/what-is-predictive-analytics/)
8. [[2008.07235] A Survey of Deep Learning for Data Caching in Edge Network (arxiv.org)](https://arxiv.org/abs/2008.07235)
9. [(PDF) Challenges and Opportunities in Edge Computing (researchgate.net)](https://www.researchgate.net/publication/307888359_Challenges_and_Opportunities_in_Edge_Computing)
10. [DeepCache | Proceedings of the 2018 Workshop on Network Meets AI & ML (acm.org)](https://dl.acm.org/doi/10.1145/3229543.3229555)