Strategic Prefetching of VoD Programs Based on ART2 driven Request Clustering

Dr. T. R. Gopalakrishnan Nair[1] and P. Jayarekha[2]

Director, Research and Industry Incubation Centre, DSI, Bangalore

trgnair@yahoo.com

Research Scholar, Dr. MGR University Dept. of ISE, BMSCE, Bangalore .

Member, Multimedia Research Group, Research Centre, DSI, Bangalore.

Jayarekha2001@yahoo.co.in

Abstract

In this paper we present a novel neural architecture to classify various types of VoD request arrival pattern using an unsupervised clustering Adaptive Resonance Theory 2 (ART2). The knowledge extracted from the ART2 clusters is used to prefetch the multimedia objects into the proxy server's cache, from the disk and prepare the system to serve the clients more efficiently before the user's arrival of the request. This approach adapts to changes in user request patterns over a period by storing the previous information. Each cluster is represented as prototype vector by generalizing the most frequently used video blocks that are accessed by all the cluster members. The simulation results of the proposed clustering and prefetching algorithm shows a significant increase in the performance of streaming server. The proposed algorithm helps the server's agent to learn user preferences and discover the information about the corresponding videos. These videos can be prefetched to the cache and identify those videos for the users who demand it.

KEYWORDS

Clustering, Predictive prefetch and neural networks.

1. INTRODUCTION

In the past few years, Multimedia applications have grown rapidly and it is evident through the exponential growth of traffic on the Internet. These applications include Video-on-demand, video authoring tools, news broadcasting, videoconferencing, digital libraries and interactive video games. The new challenges which have emerged today are related to data storage, management processing, continuous arrival of multiple requests and potentially unbounded streams that are rapid and time varying. It is generally not feasible to store the request arrival pattern in a traditional database management system in order to perform delivery operation of a video stream later. Instead, the request arrival must generally be processed in an online manner from the cache which also holds the predicatively prefetched video streams and this process assures that results can be delivered with a small start up delay for the videos accessed for the first-time.

The VoD proxy server is an important component as its function is to retrieve as many blocks of video streams as possible and send them to users. VoD proxy server is responsible for retrieving different blocks of different video streams and sending them to different users simultaneously.

DOI : 10.5121/ijist.2011.1202

This is not an easy task due to the real time commitment and the large volume of characteristics possessed by the video. Real time characteristic requires the video blocks to be retrieved from the server's disk within a deadline for continuous delivery to users. Failure to meet the deadline will result in jitters on screen during viewing. The usage of proxy server cache to store the prefetched videos can reduce the user's waiting time. In this solution the videos that are more likely to be accessed are prefetched to the browser's cache and when a request is received for one of these videos, a cache–hit occur immediately thereby avoiding start up delays. Given the previous instance, a crucial issue is to effectively predict the following requests and subsequently model the user's actions over time. If the next request can be determined accurately then the user's waiting time is reduced to zero and furthermore, no chance of user request flooding.

With the rapid development in VoD streaming services, the modern techniques to improve multimedia services has become an important research area. An important topic in learning user's request pattern is the clustering of multimedia VoD users, i.e, grouping the users into clusters based on their common interest. By analyzing the characteristics of the groups, the streaming server will understand the users better and may provide more suitable, customized services to the users. In this paper, the clustering of the users request access pattern based on their browsing activities is studied. Users with similar browsing activities are clustered or grouped into classes (clusters). A clustering algorithm takes a set of input vectors and gives as output a set of clusters and a mapping of each input vector to a cluster. Input vectors which are close to each other according to a specific similarity measure should be mapped to the same cluster. Clusters are internally represented using prototype vectors which are the vectors indicating a certain similarity.

VoD application services are made available over a computer network. It provides to watch any video at any time. One of the requirements for VoD system implementation is to have a VoD streaming server that acts as an engine to deliver videos from the server's disk to users. Video blocks should be prefetched intelligently with less latency from the disk and hence enable the service of high number of streams. However, due to real time and large volume characteristics possessed by the video, the designing of video layout is a challenging task. Real time imposes constraints on the distribution of blocks on disk and hence it decreases the number of streams being delivered to users.

In this paper, we consider the problem of clustering video streams. Thus, our goal is to maintain classes of video streams such that a selected class contains videos of more or less similar properties and attributes. The focus of the work is on delivery of video streams in real-time. Here popularity of each individual video is represented by an index, which is a measurement of popularity.

Outline of the Paper: The paper is organized into various sections as follows: Section 2 discusses about the related work in clustering and prefetching. Section 3 presents the importance of ART2. Section 4 presents the methodology used in developing the algorithm. Section 5 discusses the architecture of ART2 algorithm. Section 6 presents performance evaluation and Section 7 gives the conclusion.

2. RELATED WORK

2.1 Related work in clustering

The clustering of users based on their web access pattern is an active area of research in Web usage mining. Cooley R et al. have proposed taxonomy of Web Mining and they present various research issues. In addition, the research in web mining is centered on the extraction and applications of clustering and prefeching. Both these issues are clearly discussed by Rangarajan S K . It has been proven in this scheme a 97.78% of prediction hierarchy. Clustering of multimedia

request access pattern is defined by hierarchical clustering method to cluster the generalized session.

2.2 Related work in prefetching

Prefetching means fetching the multimedia objects much prior to the user request arrival. There are some existing prefetching techniques, but they possess some deficiency. In this the client suffers from start-up delay for those objects that are accessed for the first-time since, prefetching action is only triggered when a client starts to access that object. However, an inefficient prefetching technique causes wastage of network resources by increasing the web traffic over the network. J Yuan et al., have proposed a scheme, in which proxy servers aggressively prefetch media objects without a pattern before they are requested. They make use of servers' knowledge about access patterns to ensure the accuracy of prefetching, and have tried to minimize the prefetched data size by prefetching only the initial segments of media objects. KJ Nesbit et al., have proposed a prefetching algorithm which is based on a global history buffer that holds the most recent missing addresses in FIFO order. S K Rangarajan, et al., have proposed ART1 algorithm in which clustering and prediction has resulted in an accuracy of 97%. In this work we have proposed ART2 NN clustering algorithm for clustering user request arrival pattern. This cluster helps the server's agent in prefetching the videos into the disk, prior the user request.

3. ART2 NEURAL NETWORK

ART2 is an unsupervised neural network algorithm derived from the resonance theory. The ART networks are rather good at pattern recognizing and pattern classification. Their design allows the user to control the similarity between the patterns accepted by the same cluster]. ART2 can learn about significant new classes, yet remain stable in response to previously learned classes. Thus, it is able to meet the challenges in clustering the request arrival pattern where numerous variations are common. ART networks are configured to recognize invariant properties of a given problem domain; when presented with data pertinent to the domain, the network can categorize it on the basis of these features. This process also categorizes when distinctly different data are presented and it includes the ability to create new clusters. ART networks accommodate these requirements through interactions between different subsystems, designed to process previously encountered and unfamiliar events, respectively. We choose the ART2 neural network rather than other classifiers because capable of incrementally increasing the numbers of clusters if needed.

ART2 networks were designed to process continuous input pattern data. A special characteristic of such networks is the plasticity that allows the system to learn new concepts and at the same time retain the stability that prevents destruction of previously learned information [6].

4. METHODOLOGY

4.1 Preprocessing the web logs

The, the log data of request arrivals are represented as <client_Id,date, requested_video> format. We have selected a sample format of 50 clients requesting for 200 different videos.

4.2 Getting the Popularity Value

The popularity value is the most important parameter to get the effective prefetch operation. This popularity value may consider the long term measurement of request-frequency, which is neglected in the other algorithms. In this work the reference count value is used to get the popularity value. The reference count value is highly variable over short time scales, but this is much smoother over long time scales. This property makes the popularity value to deal with the long term measurement of request frequency.

Since the maximum and minimum value of request frequency is known during the submission of input to the ART2 system, the normalized value of the popularity can be obtained using the following transformation.

$$\delta = \frac{d - d^{min}}{d^{max} - d^{min}} \quad (1)$$

The transformed into a range between [0 ,1].

4.3 Extraction of feature vectors

For clustering, we need to extract the popularity of each video that represents the frequency of number of times the video is requested. The pattern vector maps the access frequency of each base vector element to real values. It is of the form $P = \{P_1, P_2, \ldots P_n\}$ where each P_i varies between 0 to 1.

0.8	0.2	0.2	0.1	0.3	0.4	0.5	0.1	0.3

Figure 1 Sample Pattern Vector

Figure 2 is a sample of pattern vector generated during a session.

Each pattern vector has a real value pattern of length 200.For each session we input 50 such pattern to an ART2, since we have 50 clients.

5 PROPOSED ARCHITECTURE AND ALGORITHM

Architecture of ART2 is shown in Figure 2. It is designed for processing analog as well as binary input patterns. ART2 network module includes two main parts: attentional subsystem and orienting subsystem. Attentional subsystem preprocess analog input pattern, and then choose the best matching pattern under competitive selection rule from the input pattern prototypes. Orienting subsystem carry out similarity vigilance-testing of the selective pattern prototype and trigger resonance learning and adjusting weight vectors when vigilance-testing passed, otherwise get rid of the current active node and search the other new ones. If there is no pattern prototype matching the input pattern, create a new output node to represent it. Its memory capacity can increase with the increase of learning patterns. The network allows not only off-line learning but also in an on-line learning and applying way simultaneously, that is, the learning and applying states are inseparable.

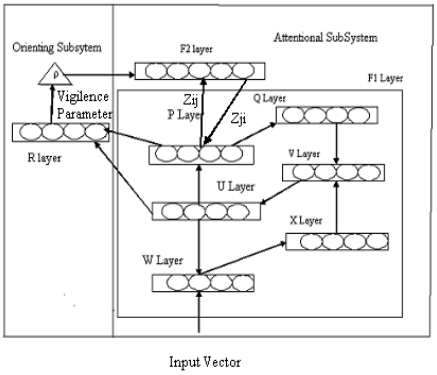

Figure 2 The ART2 Neural networks

The ART2 neural network algorithm used in this work is summarized below.

Neural network configuration:

The parameters required for ART2 formulation has been initially chosen :

Noise inhibition threshold:
$0 \le \theta \le 1$ (2)

Surveillance Parameter:
$0 \le \rho \le 1$ (3)
Error tolerance parameter ETP in the
F1 layer: $0 \le ETP \le 1$ (4)

Weight initialization of the neural
network: Top-down: $Z_{ji}(0) = 0$ (5)

Bottom-up: $Z_{ij}(0) \le \frac{1}{(1-d)\sqrt{M}}$ (6)

Operation steps:

1. Initialize the sub-layer and layer outputs with zero value and set cycle counter to one.
2. Apply an input vector I to the sub-layer *W* of the F1 layer. The output of this layer is:

$$W_i = I_i + aU_i \quad (7)$$

3. Propagate to the *X* sub-layer:

$$x_i = \frac{w_i}{e + \|W\|} \quad (8)$$

4. Calculate the *V* sub-*layer* output:

$$V_i = f(x_i) + bf(q_i) \quad (9)$$

In the first cycle the second term of (9) is zero once the value of q_i is zero.

The function $f(x)$ is given by:

$$f(x) = \begin{cases} \frac{2\theta^2}{(x^2 + \theta^2)} & 0 \le x \le \theta \\ x & x \ge \theta \end{cases} \quad (10)$$

5. Compute the *U* sub-layer output:

$$u_i = \frac{v_i}{e + \|v\|} \quad (11)$$

6. Propagate the previous output
to the *P* sub-layer: $p_i = u_i + dZ_{iJ}$ (12)

The *J* node of the F2 layer is the winner node. If F2 is inactive or if the network is in its initial configuration,

$$p_i = u_i \quad (13)$$

7. Calculate the *Q* sub-layer output:

$$q_i = \frac{p_i}{e + \|p\|} \quad (14)$$

8. Repeat steps (2) to (8) until stabilizing the values in F1 layer according to *Error* (*i*) = *U*(*i*) - U*(*i*).
If *Error* (*i*) ≤ *ETP*, the F1 layer is stable.

9. Calculate the *R* sub-layer output:

$$r_i = \frac{u_i + cp_i}{e + \|u\| + \|cP\|} \quad (15)$$

10. Determine if a reset condition is indicated. If ρ (e $\|R\|$) then, send a reset signal to F2, mark any active F2 node as not enable for competition, reduce to zero the cycle counter and return to the step (2). If there is no reset signal and the counter is one, the cycle counter is increased and passes to step (11). If there is no reset and the cycle counter is larger than one, then control passes to step (14), once the resonance was established.

11. Calculate the F2 layer input:

$$T_j = \sum_{i=1}^{M} P_i \; Z_{Ji} \quad (16)$$

12. Only the F2 winner node has non-zero output. Any node marked as non capable by a previous reset signal doesn't participate in the competition.

$$g(T_j) = \begin{cases} d & T_j = \max(T_k) \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

13. Repeat the steps (6) to (10).

14. Update the bottom-up weights of the F2 layer winner node:

$$Z_{Ji} = \frac{u_i}{1-d} \quad (18)$$

15. Update the top-down weights of the F2 layer winner node:

$$Z_{iJ} = \frac{u_i}{1-d} \qquad (19)$$

16. Remove input vector, restore inactive F2 nodes and return to the step (1) with a new input vector.

The initial values chosen for ART2 is as follows
a=10, *b*=10,*d*=0.9,*c*=0.1,*e*=0.0 θ=0.2 M = 5

Threshold value selection and methodology used

The use of several feature vectors require a special neural network, a supervised ART2 NN is used. The performance of a supervised or unsupervised ART2 NN depends on the appropriate selection of the vigilance threshold [10]. If the value of vigilance threshold is near to zero, a lot of clusters will be generated, but if it is greater, then number clusters will be generated.

6 PERFORMANCE EVALUATION

A performance analysis was done based on two parameters. (a) Hits and (b) Accuracy. Hits indicate the number of videos that are requested from the prefetched videos, and accuracy is the ratio to hits to the number of videos being prefetched. The overload of network is reduced by clustering and prefetching a community of users. It prefetches request with an accuracy as high as 97% as compared with ART1 system. In ART1 each prototype vector of a cluster represents a possibility of community of users requesting for videos in the form of binary pattern. Even a highest popular video will be represented by a value of 1. In ART2 the popularity value is normalized between [0,1].Hence the values for the input vector varies along with the popularity value.

Table 1: Result of ART 2 prefetching scheme

Number of members each cluster	Videos prefetced	Hits	Accuracy %
8	38	36	95%
6	48	46	96%
4	34	32	94%
5	31	30	96%

The results obtained have considerable improvement over ART1 which takes only the binary values as input. ART1 resulted in nearly 93 percentage accuracy prediction of prefetching .

7 SUMMARY

In recent years, there have been a number of researches in exploring novel methods and techniques to group users based on the request access pattern. In this work we have clustered and prefetched user request access pattern using ART2 neural network approach. The predictions

were done over a time series domain. The proposed system has achieved good performance with high satisfaction and applicability.

REFERENCES

[1] CARPENTER G, GROSSBERG, S , 1987, ART2: Self-Organization of Stable Category Recognition Codes for Analog Input Patterns, Applied Optics, 26: p 4916:4930.

[2] COOLEY R., MOBASHER B., SRIVATSAVA J., 1997 ,Web Mining: Information and Pattern Discovery on the World Wide Web , ICTAI'97.

[3] CHRIS TSENG H, 2007, Internet Applications with Fuzzy Logic and Neural Networks: A Survey, Journal of engineering, computing and architecture.

[4] DINKAR SITARAM, ASIT DAN, 2000, Multimedia Servers Applications,Environment, and Design, Morgan Kaufmann publishers, 2000

[5] HEINS, LUCIEN G., TAURITZ, AND, DANIEL R, 1995, Adaptive Resonance Theory (ART): An Introduction. Internal Report 95-35, Department of Computer Science, Leiden University, pages 174-185.

[6] FAUSETT, L.V. 1994, Fundamentals of Neural Networks Architectures, Algorithms, and applications. New Jersey: Prentice Hall International Inc, New Jersey, p .246-287.

[7] ISSAM DAGHER , 2006 Art networks with geometrical distances Journal of Discrete Algorithms archive Volume Issue 4 ISSN:1570-8667 Pages: 538-553

[8] JUNG J, LEE D, CHON K, 2000 Proactive Web caching with cumulative prefetching for large multimedia Data Computer Networks – Elsevier

[9] KUMAR N, JOSHI R S 2007 Data Clustering Using Artificial Neural Networks Proceedings of National Conference on Challenges & Opportunities in Information Technology (COIT-2007).

[10] NACHEV A, GANCHEV I 2003 Data Mining For Browsing Pattern Weblog Data By Art2 Neural Networks International Journal Information Theories & Applications Vol.10.

[11] NESBIT K J, SMITH J E, 2004 Data Cache Prefetching using a global history buffer, High Performance Computer Architecture.

[12] MASSEY L 2002 Determination of clustering tendency with ART neural networks Proceedings of recent advances in soft-computing (RASC02) .

[13] RANGARAJAN S K , PHOHA V V , BALAGANI K, SELMIC R R 2008 Web user caching and its application to prefetching using ART neural networks IEEE Internet Computing, Data & Knowledge Engineering, Vol.65 No.3, p.512-543

[14] SANTOSH K, VIR V, KIRAN S, SELMIC R, IYENGAR SS, 2004 Adaptive Neural Network clustering of Web users IEEE Computer..

[15] TAE-UK CHOI, YOUNG-JU KIM,1999 "A prefetching scheme based on the analysis of user access patterns in news-on-demand system", International Multimedia Conference Proceedings of the seventh ACM international conference on Multimedia (Part 1) Orlando, Florida, United States, ISBN:1-58113-151-8 Pages: 145 - 148 .

[16] VENKETESH P, VENKATESAN R 2009 A Survey on Applications of NeuralNetworks and Evolutionary Techniques Web Caching IETE Technical review.

[17] YUAN J, SUN Q, RAHARDJA S, 2007 A More Aggressive Prefetching Scheme for streaming Media Delivery over the Internet Proceedings of SPIE

[18] XIAOBO ZHOU " A video replacement policy based on revenue to cost ratio in a multicast tv-anytime system.

Authors

T.R. Gopalakrishnan Nair holds M.Tech. (IISc, Bangalore) and Ph.D. degree in Computer Science. He has three decades experience in Computer Science and Engineering through research, industry and education. He has published several papers and holds patents in multi domain. He won the PARAM Award for technology innovation. Currently, he is the Director of Research and Industry in Dayananda Sagar Institutions, Bangalore, India.

P Jayarekha holds M.Tech (VTU Belgaum) in Computer Science securing second rank . She has one and half decades experience in teaching field. She has published many papers. Currently, she is working as a teaching faculty in the department of Information Science and Engineering at BMS College of Engineering, Bangalore, India.