Investigation on the Content Popularity Distribution under K-Transformation in Streaming Applications

Zongkai Yang, Tai Wang, Xu Du, Wei Liu, and Jiang Yu
Dept. of Electronics and Information Engineering, Huazhong Univerisity of Science and Technology, China
{zkyang, wliu}@public.wh.hb.cn, wang.tai@163.com, {duxu, frankyu}@263.net

***Abstract*— An accurate understanding on the content popularity distribution of media objects in streaming applications is required when designing efficient caching algorithms for the streaming media. However, there is no empirical model to describe this distribution. Our investigation shows that the content popularity of a hot streaming media file exhibits a Zipf-like distribution under K-Transformation, after analyzing 250 hot clips in 2 long-term traces. A specific function is proposed to describe this phenomenon. This result builds a foundation for designing intelligent caching algorithms.**

***Index Terms*— Content Popularity Distribution, Proxy Caching, Streaming Media**

I. Introduction

As the broadband Internet is becoming common and the multimedia technology has been widely used, the streaming media service consumes much more bandwidth. An appealing way to reduce the backbone bandwidth consumption is the proxy caching, which is originally used to improve the quality of web content delivery. Those proxies deployed at the edge of the networks can reduce clients' start-up delay and backbone bandwidth consumption significantly by caching hot objects. However, different from web objects, streaming media files are in usually huge size. A one-hour MEPG-1 video file is almost 700MByte. It is not practical to cache such a large file as a whole in the proxy. Due to this reason, designing efficient caching algorithms attracts people a lot.

At present, most streaming media caching algorithms summarized in [1] divide a media file into segments to improve the disk space utility. The segments are the objects of the proxy operation such as prefetching and replacing. Generally speaking, the larger the probability of a hot segment being cached, the less the backbone bandwidth consumption will be achieved. As a result, the segment popularity is usually used as the main metric to determine whether a certain segment should be cached [2]-[5]. A key obstacle to design such an algorithm is the lack of an accurate understanding on the content popularity of a streaming media file. Its importance to the algorithm's performance is presented specifically in Section II.

Several studies [6]-[10] indicate that the content popularity of an entertainment media clip decreases from the beginning to the end because of the *early abort* cases. Nevertheless, there is no empirical model to describe this phenomenon specifically.

Existing streaming proxy caching algorithms usually take the volume of one segment whose playback time is 1 second as the unit. Therefore, motivated by this, we investigate the popularity distribution of each second in a video clip, which is called the content popularity, different from the external popularity that only counts the request times of a certain clip.

The streaming media online takes the form of video and audio clips from news, sports, entertainment and educational sites. We get 2 long-term traces provided by a considerably representative web site, Yangtse Interactive Media Network(www.cjmedia.com.cn), offering streaming media entertainment services. We analyze the content popularity of 250 hot video clips under K-Transformation. The results show that the hot clips' content popularity distribution follows a Zipf-like distribution very well. This result builds a foundation for designing new efficient streaming proxy caching algorithms.

The remainder of the paper is organized as follows. We review related work in Section II. Section III observes the content popularity distribution in normal methods. Section IV focuses on the temporal locality, *i.e.* nearby segments response similar requests. Section IV introduces a coordinates transformation with the parameter k to restrain the temporal correlation between nearby segments, which is called K-Transformation. Section V shows that with the help of K-Transformation, the new coordinates of a segment's popularity and its position exhibit a Zipf-like distribution and this phenomenon exists in most hot clips. Finally, Section VI presents the conclusion and future work.

II. Related Work

As the Internet's fast growing up, Breslau *et al.*[11] discovered several evidences that support the common existence of the Zipf-like distribution in web objects such as text, graph, and small video or audio files, *i.e.* the access frequency of the r-th most popular file is proportional to $1/r^{\alpha}$. A larger α implies more sessions are concentrated on the most popular files. If the frequencies of files and the corresponding popularity ranks are plotted on a log-log scale, a Zipf-like distribution can be fitted by a straight line. The access frequency of an object is usually called its popularity. Based on the object's popularity, a number of web caching algorithms were proposed [12].

The work in this paper was supported by the National Natural Science Foundation of China (No.60302004) and also partially supported by the Australian Research Council grant (LX0240468).

Almeida *et al.* [7] found that if the whole media file is regarded as the web object, there is a similar distribution as well. However, as the example mentioned in Section I, people can not simply clone the web caching algorithm to improve the streaming proxy caching performance due to the huge size of media files. As a result, researchers propose segment-based caching algorithms to improve the disk space utility.

As the understanding of the content popularity is becoming more and more accurate, the streaming caching algorithms which are proposed one after the other perform better and better.

Subhabrta *et al.* proposed the prefix caching algorithm, which prefetches a small portion at the beginning of a clip to reduce the clients' perceived start-up delay [13]. Soam *et al.* discovered that more than 50% sessions end before a clip's 5% position [6]. This *early abort* phenomenon explains why the prefix caching is efficient to improve a proxy's byte hit rate and reduce the backbone bandwidth consumption.

A lot of studies demonstrate that the content popularity of an entertainment clip decreases from the beginning to the end since the *early abort* cases are common, *e.g.* only 29% sessions finish a whole clip [8]. Based on this phenomenon, Wu *et al.* [2] set the reciprocal of a segment's position in the clip as its popularity. This assumption is a reasonable estimation on the hit probability of a segment, which becomes one of the explanations that why the byte hit rate of their algorithm is 15% higher than the one of the prefix caching algorithm on average. To estimate a segment's popularity more precisely, Songqing *et al.* [3] created a more sophisticated caching cost function, achieving 30% higher than the former in byte hit rate on average.

From the review above, it is convincible that the proper choice of the content popularity model is important to improve the performance of the streaming proxy caching algorithm. However, most of the researchers concentrate their attentions on sessions when they analyze the streaming server workload, such as the playback time distribution or the transferred byte distribution and so on. There is few mathematic description on the practical content popularity of a clip, although Chae *et al.* assumed that it is a negative exponential distribution [4].

III. Content Popularity of Streaming Media

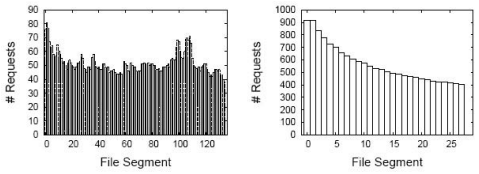
The trace sources are 2 long-term traces provided by Yangtse Interactive Media Network(www.cjmedia.com.cn). Yangtse Interactive Media Network is a powerful news media web site in Wu Han city and also a member of the *City League* which only consists of mainstream news medium one city each, approved by State Council Information Office. It also offers streaming media services. The site hosts diverse video coverage of movies and teleplays made in China, Japan, Korea, Europe, and American. The clips include action films, affectional films and many other kinds of entertainment videos. Table I briefly summarizes 2 long-term traces, named Trace I and Trace II, respectively.

Section III and Section IV will present our investigation on Trace I. The same investigation procedure is also performed on Trace II.

TABLE I
STATISTICS SUMMARY FOR THE SITE

	Trace I	Trace II
Log Duration	91 days	210 days
Total Sessions	43,262	92,059
Unique Clips	3,542	6,209
Date Range	04/01/04-06/30/04	04/01/04-10/27/04

Up to now, the most similar work to ours is C. Costa *et al.*'s. Their work [10] implies that the trend of the content popularity distribution of a certain clip is impacted by the clip's category, as shown in Fig. 1 from [10].



(a) Most Popular Files (Educational)

(b) Less Popular Files / Entertainment Files

Fig. 1. File Segment Access Frequencies [10]

They discover that the most popular content can appear not only in the beginning, but also in other portions of the clip when the clip is for education. This phenomenon can be explained by the fact that some students will jump over the content that they are already familiar with, to the content that they can not catch up. There they will jump back a little and playback the content again and again until they can follow. This usually happens in the listening practise in a foreign language lesson. The motivation of mastering new knowledge forces the students to do so. This trend of the content popularity distribution is the *Random* mode in [5].

On the contrary, if the clip is for entertainment, the behavior of the audience are quite different. When they feel the clip is boring or unreadable or even in unbearable quality of service, they just stop resuming right there and leave for a new clip. That is the *early abort* mode adopted by many researchers [2][3][4][13] when they model the content popularity distribution. The monotonically decreasing curve shown in Fig.1 implies that when the audience are interested in some content, they would like to choose to go on without any interrupt, expecting another surprise ahead, rather than pause there and review the same content again and again.

The reasoning above is also supported by the different numbers of VCR actions per session in different kinds(entertainment or education) of clips presented in [10]. In fact, approximately 22%(28%) of the sessions to 20-30(30-40) minute files have 10 or more interactive requests, observed in the eTeach trace in [10], while 2 or more interactive requests were observed in less than 5% and 2% of all sessions in the RADIO/UOL and ISP/Audio workload in [10].

The 2 long-term traces that we investigate in this paper are collected from the Helix Universal Servers, providing

entertainment videos. The former of Helix Universal Server is Real Server G2. The traces record the clients' VCR-like operations such as jump forward/backward or pause and so on. Following the instructions in the administrator guide [14], we can know at which second the client interacts with the video clip, and what kind of VCR-like operations the action is.

It is observed that the number of VCR-like operations per session is around 5 times, *i.e.*, the number of jump or pause actions per session is less than 3 times, excluding resume or stop actions. Comparing with educational clips [10], and considering the reason stated above, clients' interactions such as jump or pause have very limited influence on the content popularity distribution of an entertainment clip. Hence, it is reasonable to ignore users' interactive behavior occurring within the session except the resume action at the beginning and the stop action at the end of the session. On the other word, we investigate the content popularity on the 2 assumptions: 1) users view a clip from the beginning of the clip, and can possibly end the session at any position of the clip; 2) there is no action after the resume action at the beginning of the clip or before the stop action at the end of the session.

The parameters used in investigating a clip's content popularity are listed in Table II.

TABLE II
PARAMETERS FOR INVESTIGATING

V	the clip's name, used as its ID
L	the clip V's length(in sec.)
S	the number of the clip V's sessions
j	the sequence number of a one-second-long segment in the clip V, $j = 1, 2, ...L$.
i	the sequence number of a session of the clip V, $i = 1, 2, ...S$.
A	A is a vector. $A(j)$ is the viewed times of the clip V's j-th second. Define: $A(L+1) = 0$.
R_i	R_i is a vector. R_i records the clip V's each second's viewed status in the clip V's i-th session. $R_i(j)$ is defined in (1).
α	the concentration degree in the Zipf-like distribution
C	the coefficient in the Zipf-like distribution
(x, y)	the x-th second has been viewed y times.
k_x	the squeeze factor in K-Transformation along X-axis
k_y	the squeeze factor in K-Transformation along Y-axis
(x', y')	the new coordinates of (x, y) after K-Transformation

$$R_i(j) = \begin{cases} 1 & (if\ the\ jth\ second\ is\ viewed) \\ 0 & (if\ the\ jth\ second\ is\ not\ viewed) \end{cases} \quad (1)$$

According to the definition of A in Table II and (1),

$$A = \sum_{i=1}^{S} R_i. \quad (2)$$

Based on the assumptions 1) and 2), in the clip V's i-th session, the client views every second before he stops at the position of the $d(i)$-th second. Thus,

$$R_i = [\underbrace{1, 1, 1...1, 1, 1,}_{total\ of\ d(i)\ 1}\ \underbrace{0, 0, 0...0, 0, 0}_{total\ of\ (L-d(i))\ 0}\ ] \quad (3)$$

Then, $R_i(j) \geq R_i(j+1)$. Thus,

$$A(j) = \sum_{i=1}^{S} R_i(j) \geq \sum_{i=1}^{S} R_i(j+1) = A(j+1). \quad (4)$$

Following (2) and (3), we figure out the top 1 clip's content popularity shown in Fig. 2. Equation (4) demonstrates the clip's content popularity is not increasing from the beginning to the end based on the 2 assumptions stated above. Therefore, the j-th second's popularity is ranked at j in all the seconds' popularities, as shown in Fig. 2. Considering that the Zipf-like distribution describes the relationship between an object's popularity and its rank, we attempt to fit the content popularity curve in Fig. 2 with a Zipf-like distribution. However, the top 1 clip's content popularity exhibits a circular curve in log-log scale as shown in Fig. 3.

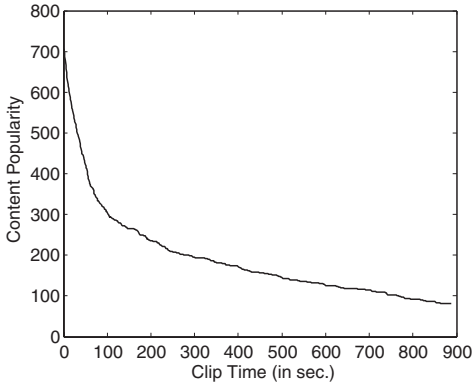


Fig. 2. Top 1 Clip's Content Popularity Distribution

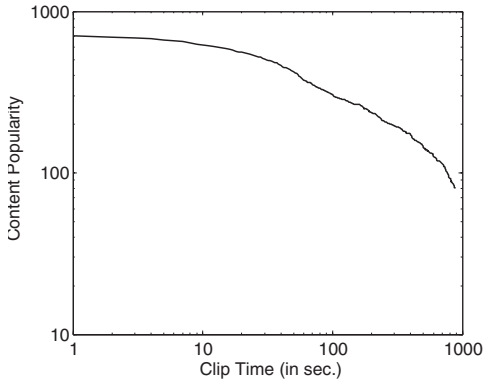


Fig. 3. Top 1 Clip's Content Popularity Distribution in log-log Scale

Obviously, the top 1 clip's content popularity does not follow the traditional Zipf-like distribution. And it is difficult to describe the circular curve in an easy form, resulting in an inaccurate estimation on the content popularity.

IV. Content Popularity's K-Transformation

One possible explanation on why the content popularity does not follow a Zipf-like distribution is that the trace duration is so long that there is little differentiation in the

frequencies of the beginning or ending part of the hot clips, as shown in Fig. 2. This inspires us to gather nearby 1-second-long segments into a chunk to enlarge the differentiation in each chunk's popularity. The second motivation is that clients usually decide to stop the session after they view a meaningful group of pictures, which lasts for a few seconds. Another reason is that if we regard a whole clip as a series of small chunks, the chunks may exhibit a Zipf-like distribution where K-Transformation is successful in modeling the external popularity of the clips [15].

It is assumed that each chunk contains k_x(see Table II) continuous 1-second-long segments. The x'-th most popular chunk's popularity is represented by the $((x'-1)k_x+1)$-th second's popularity, which is the head segment of the x'-th chunk, since the segments inside a chunk are suppose to have similar popularities. Assuming that x is the position of the head segment in the clip and x' is the corresponding chunk rank, x' and x satisfy (5). Other chunk ranks are transformed to float numbers evenly distributed among integral ranks.

$$x' = \frac{x + k_x - 1}{k_x} \tag{5}$$

Based on the similar consideration, we should not only gather the segments along the dimension of clip time, but also gather the original content popularities along the dimension of content popularity. Hence, y' and y satisfy (6).

$$y' = \frac{y + k_y - 1}{k_y} \tag{6}$$

We call the coordinates transformation shown in (5) and (6) K-Transformation. In fact, k_x and k_y are the squeeze factors along X-axis and Y-axis in Fig. 2 and Fig. 3, respectively.

We perform K-Transformation on the data in Fig. 2, assigning both k_x and k_y with 40. The new curve in log-log scale is shown in Fig. 4.

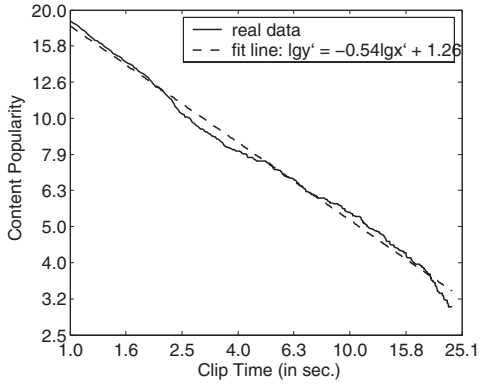


Fig. 4. Top 1 Clip's Content Popularity Distribution in log-log Scale under K-Transformation

The dotted straight line in Fig. 4 is described by (7).

$$\lg y' = -\alpha \lg x' + 10^C \tag{7}$$

Comparing with Fig. 3, Fig. 4 shows that under K-Transformation, the top 1 clip's content popularity distribution follows a Zipf-like distribution very well. It also shows that the importance of K-Transformation in the investigation on the content popularity of a video clip. We use R^2 to measure the goodness of a regression. If $R^2 = 1$, the model is perfect. A larger R^2 implies a more accurate model [16]. For Fig. 4, R^2 is 0.987.

From (5), (6), and (7), we can get (8).

$$\frac{y + k_y - 1}{k_y} = \frac{C}{(\frac{x+k_x-1}{k_x})^\alpha} \tag{8}$$

Section V verifies that this distribution exists widely in other hot clips under K-Transformation.

V. K-Transformation's Validation in Describing the Content Popularity Distribution

In this section, we will follow (2), (5) and (6) to investigate the top 100 hot clips' content popularity distributions in Trace I. Although they are only 2.8% of 3542 clips, they take 23.6% requests and 20.5% view duration of all. Therefore, these hot clips are very important for describing the content popularity pattern. The result shows that among top 100 clips, there are 35 clips' R^2 no less than 0.95, and there are 67 clips' R^2 no less than 0.90. It is also noticed that only 11 clips' R^2 are less than 0.90 in top 50 clips while this amount is doubled from the top 51 to the top 100 clip.

The investigation result of Trace II, whose duration is 2.3 times of Trace I, shows that among the top 150 clips in 210 days, only 4% of the clips follow a Zipf-like distribution without K-Transformation whose R^2 are no less than 0.95. Under K-Transformation, the number of clips whose R^2 are no less than 0.95 is up to 90(60% of the 150 clips). This result not only shows again that K-Transformation is helpful to investigate on the content popularity distribution of hot video clips, but also verifies that the reason that why K-Transformation is failed for some clips is the lack of sufficient request sessions.

And we also find that the change of k_x's or k_y's value has a little but not much effect on the improvement of R^2. Take Trace II as an example. In Trace II, we call the video whose R^2 is larger than 0.95 the *successful* clip. It seems that the number of successful clips will increase when k_y increases while decrease when k_x increases. And the number of successful clips seems to be constrained by an upper limit since it won't change after it reaches 91 when k_x is 10 and k_y changes from 340 to 500. We list the number of successful clips when k_x is 10 and k_y is 50, 100, 200, 300, 400, 500 in Table III .

TABLE III
Influence on Successful Clip Numbers from k_y in Trace II

k_y	Successful Clips
50	72
100	80
200	88
300	90
400	91
500	91

The parameters of the Zipf K-Transformation of the Trace I and Trace II are listed in Table IV. The specific function describing the content popularity distribution of a hot entertainment clip is shown in (8).

TABLE IV
PARAMETERS OF THE ZIPF K-TRANSFORMATION

Trace	R^2	successful clips	hot clips	mean α	k_x	k_y
I	≥ 0.90	67	100	0.23	40	40
II	≥ 0.95	90	150	0.05	10	300

VI. CONCLUSION

An accurate understanding on the content popularity distribution of a streaming media file is helpful to design efficient caching algorithms. We investigate this distribution using 2 long-term traces. Our investigation shows that K-Transformation is a useful tool to capture this distribution with a Zipf-like distribution. This conclusion can be combined with the external object's popularity distribution by one general distribution, which builds a foundation for designing intelligent caching algorithms, take [17] for example.

And we do notice the fact that the content popularity distribution of educational clips can show different curves as presented in Fig. 1. We will go on to look for new traces provided from e-learning web sites to investigate.

ACKNOWLEDGMENT

The authors would like to acknowledge Yangtse Interactive Media Network for providing 2 long-term traces. They would also like to thank the anonymous reviewers for constructive comments which helped to improve the content and presentation of the paper.

REFERENCES

[1] J. Liu and J. Xu, "Proxy caching for media streaming over the internet," *IEEE Communications Magazine*, vol.42, no.4, pp.88-94, Aug. 2004.

[2] K. Wu, P.S. Yu, and J.L. Wolf, "Segmentation of multimedia streams for proxy caching," in *IEEE Transactions on Multimedia*, vol.6, no.5 pp.770-780, Oct.2004.

[3] S. Chen, B. Shen, S. Wee, and X. Zhang, "Adaptive and lazy segmentation based proxy caching for streaming media delivery," in *Proc. of 13th Int. Workshop on NOSSDAV.*, Monterey, CA, Jun.2003, pp.22-31.

[4] Y. Chae, K. Go, M.M. Buddhikot, S. Suri, and E.W. Zegura, "Silo, rainbow and caching token: schemes for scalable tolerant streaming caching," *IEEE Journal on Selected Areas in Communications*, vol.20, no.7, pp.1328-1344, 2002.

[5] W. Liu, C.T. Chou, Z.K. Yang, and X. Du, "Popularity-wise Proxy Caching for Interactive Streaming Media," *Proc. 29th IEEE LCN*, pp.250-257, Nov. 2004.

[6] S. Acharya, B. Smith, and P. Parnes, "Characterizing user access to videos on the world wide web," in *Proc. SPIE/ACM Multimedia Computing and Networking Conf.*, San Jose, CA, Jan.2000, pp.130-141.

[7] J.M. Almeida, J. Krueger, D.L. Eager, and M.K.Vernon, "Analysis of educational media server workloads," in *Proc. 11th Int. Workshop on NOSSDAV.*, Port Jefferson, NY, Jun.2001, pp.21-30.

[8] L. Cherkasova and M. Gupta, "Characterzing locality, Evolution, and life span of accesses in engerprise media sever workloads," in *Proc. 12th Int. Workshop on NOSSDAV.*, Miami, FL, May 2002, pp.33-42.

[9] W. Tang, Y. Fu, L. Cherkasova, and A. Vahadat, "A synthetic streaming media service workload generator," in *Proc. 13th Int. Workshop on NOSSDAV.*, Monterey, CA, Jun.2003, pp.12-21.

[10] C. Costa, I. Cunha, A. Borges, C. Ramos, M. Rocha, J. A, and B. Ribeiro-Neto, "Analyzing client interactivity in streaming media," in *Proc. 13th Int. World Wide Web Conf.*, New York, NY, May 2004, pp.534-543.

[11] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Scott, "Web caching and Zipf-like distributions: evidence and implications," in *Proc. IEEE INFOCOM.*, Mar.1999, pp.126-134.

[12] J. Wang, "A survey of web caching schemes for the internet," *ACM SIGCOMM Computer Communication Review*, vol.29, no.5, pp.36-46, 1999.

[13] S. Sen, J. Rexford, and D. Towsley, "Proxy prefix caching for multimedia streams," in *Proc. IEEE INFOCOM.*, Mar.1999, pp.1310-1319.

[14] Helix Universal Server Administrator Guide. RealNetworks, Inc. [Online]. Availble: http://docs.real.com/docs/HelixServer9.pdf.

[15] L. Cherkasova, and M. Gupta, "Analysis of enterprise media server workloads: access patterns, locality, content evolution, and rates of change," *IEEE/ACM Transactions on Networking*, vol.12, no.5, pp.781-794, Oct. 2004.

[16] R. Jain. *The art of computer systems performance analysis: technique for experimental design, measurement, simulation and modeling.* John Wiley & Sons, pp.221-240, 1992.

[17] J. Yu, C.T. Chou, X. Du, T. Wang, " Internal popularity of streaming video and its implication on caching", *Technical Report*, May, 2005.