Location Based Content Recommendation for CASoRT System

Wan Dong*, Xiaofeng Zhong*, Naijia Liu†, Pengzhi Xu‡ and Jing Wang*
*Department of Electrical Engineering, Tsinghua University, Beijing, China 100084
Email: {dongw11@mails.tsinghua.edu.cn, zhongxf@tsinghua.edu.cn, wangj@tsinghua.edu.cn}
†Computer and Information Management Center, Tsinghua University, Beijing, China 100084
Email: lnj@cic.tsinghua.edu.cn
‡Department of Computer Science, Tsinghua University, Beijing, China 100084
Email: xpz03@163.com

Abstract—The current wireless communication network is a content-careless network, in which every request needs a new transmission, even those requests imply the same content. However, the Long-Tail distribution of users’ interests caused by the converging behavior, determines that the redundant traffic, which is the re-transmission bearing same content, is rather enormous. Therefore the content-careless scheme of network transmission introduces large energy waste. In our previous work, we put forward a novel transmission scheme CASoRT to mitigate such energy waste by broadcasting those hot contents. After the contents are broadcast, the terminals in coverage could cache those contents in local memory and load those contents when request. Thus, the massive redundant transmission of the very content is relieved by one broadcast transmission. In order to exploit the advantage of broadcast, the proper recommender scheme, deciding which contents should be broadcast, needs to be optimized. In this paper, we present certain perspective in analyzing the log of campus network in Tsinghua University. It shows that utilizing users converging character could improve the efficiency of CASoRT with analysis. Finally, location based content recommendation scheme is proposed, verified by simulation.

I. INTRODUCTION

The facts that people shares common interest within certain time duration and geographic area, in other words the converging character of human behavior, are well-acknowledged. In network research area, the Long-Tail distribution of human behavior, induced by the converging character of human behavior, has come to a consensus. In network operation, the most notable phenomenon caused by such human nature is that large percentage of people would frequently request and browse the contents within a small but hot group, such as the latest news and recent movies. However, each request, with regardless to wether such content has been transferred before, needs a new transmission. The redundancy is implied under such service scheme as the data packages, bearing the same content, transmit in network for many times. Moreover the percentage of such redundant transmission in network, by the Long-Tail distribution theory, is not trivial but rather enormous. With each repeated request on same content would also consume almost same amount of power, the traditional service mode introduce another kind of energy waste, the content-careless waste, in parallel with energy wastes in physical transmission. Such kind of energy waste takes considerable percentage in energy consumption in current network.

In wireless network, the broadcasting nature of transmission could be utilized to mitigate such shortcoming in traditional network. In wireless network, end-to-end transmission could be listened by all users in coverage, if not filtered. Thus, the broadcasting like transmission makes that one time broadcast would not consume too much additional energy, compare to end-to-end transmission. If their device could memorize the broadcasted content from network in local memory and load from local when such content is request, the transmission could be replaced by reading from local memory and the re-transmission in network could be alleviated. With exploiting users’ converging properties, the broadcasting scheme in wireless scenario could be more efficient. By broadcasting the hot content to cover proper range, high possibility, that the content would be accessed in future, would be achieved, with rational broadcast radius and energy consumption.

With this spirt, our previous work[6] proposes the novel CASoRT System(Content Aware Soft Real Time System) to alleviate such waste and achieve better energy efficiency. Additionally, in[3], the broadcast scheme that achieves sub-optimal energy efficiency is proposed, with the assumption that the model for predicting the possibility, that users would access certain file, is provided. However, the broadcast scheme needs more than a proper broadcast protocol, but a highly efficient recommender system for the specific scenario. The traditional recommender system would perform less effective in CASoRT system, as the fundamental presumption on network principle is different. Therefore, for higher energy efficiency to be achieved, more suitable content recommendation scheme for CASoRT should be proposed.

The paper is organized as follows: section II presents the related work. Section III presents methodology of data collection and raw data analysis, in order to prove the representativeness of the data. Section IV presents the analysis result on converging in location and content in request, and then proposes the location based recommender scheme.
II. RELATED WORK

The solution to reduce such repeated traffic in network have already been invented in network field. The incipient attempt is IGMP (Internet Group Management Protocol) protocol [2]. By caching packages in IGMP router and re-broadcasting them to users in sub-broadcasting group in network layer, the protocol could reduce redundant package transmission between Video server and IGMP router. In recent, with the idea of cross-layer optimization, [5] proposes the scheme in cloud service that propagate the locally interesting content to the boarder gateway of cloud that nears the corresponding user group. Via such scheme, the traffic in back-bone network of the cloud could be lessened. However those methods are less effective in our scenario, as the bottle neck is the wireless transmission, rather than the wired back-bone network behind base station.

The scheme of recommending proper content to certain group people, the recommender system, has been explored in computer science region for long time and some of the theories have been applied in existing system. [4]summarizes the framework of collaborative-filtering method and content-base filtering method. The collaborative-filtering method emphasis on the relationships between people. By analyzing users’ trace of history information, the system could decide to whom this user is akin. With the creed that people who shares similar interest in past would shares interest in future, the system would recommend the items from the history trace of related users. The content-based filtering method is derived from data mining technology. Via analyzing the users’ history records in statistic, the system could classify the users based on the content they are interested in. Then, the system would recommend the content based on its class, that is recommending the content to the users who belongs to the very class. In recent years, the hybrid method that absorbs the advantages of both methods is invented[1].

Most of the recommender systems focus on personalization, providing suitable content to specific user. Little work has involved with recommending the generally interesting content to massive group people, which is the intrinsic requirement of the content recommendation scheme for CASoRT system. Also in research region of traditional network, the geographic distribution is often ignored. Most research is based on the presumption that the users are distributed sparsely in geometry distribution is often ignored. Most research is based on the presumption that the users are distributed sparsely in geometry because the more influential factor of network performance is topological network distribution. Comparing to the traditional method of recommendation, our research weighs on the converging character of users’ geographic and interest distribution, for those are important factors influencing the efficiency of broadcasting system.

III. DATA COLLECTION AND RAW ACCESS LOG DATA ANALYSIS

A. Data Collection

Our log data is obtained in the gateway server of a cloud data service site operated in Tsinghua campus network, called Meepo (thu.meepo.org). Meepo could only be accessed via campus network and it is heavily used by Tsinghua students. The structure of Meepo is depicted in Figure 1. All of the accesses to files stored in Meepo cloud should pass the gateway server and each operation on files would leave record on access log in gateway server. The log data is obtained from Meepo site network gateway. By cross query the imported log data in database, the exact location, time, content and user id of each access could be extracted.

The files stored in Meepo consists as follows: 1.public files, such as teaching video or public document, are uploaded by site administrator periodically and could be accessed by all users. 2.private files, the personal data of individual, is uploaded by students and for those files Meepo is served as ubiquitously accessing storage in campus. Our analysis focuses on the access of public files in Meepo, which could reflect the access pattern of public data in real network, at least in certain scenario. As Wi-Fi access points are deployed and could cover whole Tsinghua campus, the users’ access pattern extracted in the log data could reflect mobility pattern like cellular network, and thus this work could represent some aspect of the character of users’ behavior in mobile Internet service. Additionally, we believe that the most essential influence factor of users’ behavior is the innermost needs of humanity and the impulsion of human instinct. Thus, the users’ behavior and the implied behavior model would keep relative constant with regard to the variance of support network media. So even if our data, which is sampled from campus network, has certain confinement than real cellular network, the model obtained, which focuses on the behavior behind the superficial phenomenon in access log, could be representative and has certain value in research area.

Amongst all files stored in Meepo storage bank, we choose the video files as the representative type of file to research. The movie files have higher purity in its bearing content than any other file types, such as text and audio files. Video files almost represent a video clip with certain subject, while the substance of text and audio files range far more extent. It makes filtering out the log noise, which is the log data created not by actual access, easier than other file types, considering the log data recording mechanism of Meepo and the character of video

![Fig. 1. Site Structure of Meepo](image-url)
playback. Also we obtained the meta data of each movie file, including size, file type, create time while extracting detail from other file type is difficult. Moreover, analyzing the access pattern for video files is most meaningful, since the size of video files is much larger in average than other file types. It means that broadcasting video files would obtain better performance.

B. Raw Data Analysis

Tsinghua campus, which is a $2\text{km} \times 2\text{km}$ area in approximate, covered by the campus wireless network. There are about 17.3 thousand Meepo users in aggregate, which mainly consist of students. There were 77.74TB data stored on Meepo and the size is still growing. We sampled the log data from 14th November, 2012 to 16th December 2012, total 33 days.

The number of distinct active users is shown in fig 2. During that period, there are 7573 users who had accessed to Meepo in total and about 1300, about 17.2% users would access to Meepo site per day in average. Each user would access to Meepo 5.68 times in each day in average, which means that heavy users who accessed to Meepo in one day would tend to have multiple request. It could be inferred from the facts above that quite a lot of users would access Meepo frequently and their access have clearly fixed pattern, which is the target for analysis. Periodical pattern could be observed obviously during the period, about 8 to 10 days in each period. The active character of content may yield such pattern, and it would be further discussed in next section.

The distribution of users' active days is shown in fig 3. In fig 3, there are about 1000 users would access to Meepo more than 16 days and the decay is relative slow with respect to number of days, which illustrates that people who have custom to access to Meepo take relative less proportion with respect to those users who would visit sporadically. However, the traffic induced by those heavy users in total would take considerable proportion, for that, in multiple days, they would introduce massive visit each day. Moreover, as the access pattern is different between those two group of users, distinct cluster in access pattern may exist. In next section, we discuss in detail about such clustering phenomenon.

IV. CLUSTER ANALYSIS ON USERS’ BEHAVIOR AND LOCATION BASED CONTENT RECOMMENDATION

A. Analysis of Geometrical Convergence

By intuitively thinking, users would pursue most recent video and news, which is geometrically independent. However, by the nature of non-uniform distribution of population, people would converge at specific places and therefore it makes the access pattern of those content geometrically converging. Also with the drift of distribution of population over time, the detailed character would also change with time. The location related contents, such as local news, would only be appealing to people dwelling in limited area. The inherent makes such content converging strongly in geographic. Moreover, as the growing up of social network, more and more user created content is becoming a important part in network. Such content would almost transfer to limited region as the confinement of social attribute of humanity. Thus, it causes the phenomenon that people requesting such content would locate in small region.

As the restriction of deployed area of campus network and the limitation of the structure of Meepo site, it is hard to prove the second and third assumptions above. In this chapter, we use the log data sampled in Tsinghua University to prove the first hypothesis. Fig 4 and fig 5 represents the distribution of users who requested the top 5 hot video content in Meepo. The access converge in dormitory area (North of campus), administrative office (West of campus) and two R&D building (East south of campus) in daytime and almost all request happened at dormitory area at night. Such phenomenon strongly support the first hypothesis in last paragraph. Those contents are all location independent, while the access pattern in location reveals imbalance in geographic distribution. This could be explained by the limitation of range of activity on human in society. People would only have chance to be exposed to specific information within limited area. Also, bounded by regulation on behavior, people would tend to stay
at fixed location, such as home and office, for relative long time, and then commute to other places with the alternation of social duty.

Fig 6 represents CDF of the average number of places users would access to Meepo site. Most people, over 80% in total, would only have two places in average that accessed in Meepo. From fig 6, and the argument above, the conclusion that, users have strongly fixed access pattern in location, could be made. That the restriction of social activity makes people could only stay in two or three fixed places everyday could be a reasonable explanation.

### B. Analysis of Access Converging in Content

Fig 7 describes the result of k-means cluster analysis on access number of each file in time domain and the number of files in each cluster is described in table 1. If the parameter of cluster number is more than 9, as set in our analysis, new cluster would appear near the existing cluster center in feature space. Therefore, more accessing pattern of files in Meepo site in time domain would not be discovered by increasing the number of cluster. Also, the number of files whose access number is over 200 during 33 days is compatible with the number of files which in the clusters that has distinct peak in access number during 33 days.

The files in Meepo site could be distinguished in two groups: most files would only be scarcely visited by users in 33 days, while minor number of files would get concentration of access in a time period. The persistent time of popularity of hot file only keeps about 10 days after the first appearance of concentrated visit and then the number of access would attenuate to low level fast. Then, such hot files would degenerate to the pattern, like most of files, that seldom visited by users and another peak in access number would not reappear again. It could be deduced that the general trend of all files should be the same, with the only differences among all files are that the access number in hot period and the time that the file comes into hot period. The about 10 days active period echoes with the periodical pattern depicted in fig 2. The periodical pattern in overall access behavior also evidences the reasonableness of such hot period pattern.
C. Location Based Content Recommendation

Our content recommendation scheme for CASoRT system is grounded on the clustering property of access pattern in geological and content dimension. The inherent of recommender system is to discovery the potential connection between users’ access behavior and use the relevance as principle to filter content and push the suitable content to related users. In our work, we find the strong connectivity between location and the clustering property of interest, and such connection could be utilized as such guideline for filtering files. The recommendation scheme is more suitable to CASoRT system than the traditional recommender system. By recommending the file bearing those hot content to CASoRT system for broadcasting to related location, massive re-transmission could be reduced, with the high concentration in interest in certain location, with controllable broadcasting energy. The location-related scheme could be utilized by CASoRT system to evaluate and balance the potential energy reservation and broadcasting energy, and then the CASoRT system could further optimize the energy efficiency by setting the most proper broadcasting radius with estimated energy consumption and reservation in broadcasting scheme level, as described in [3].

The concentration of access in content and fast decay of popularity also makes our recommendation scheme suitable for the application scenario of CASoRT system. The access pattern that converging in few content makes it possible for recordation scheme that focus on those hot content, while neglecting most inactive files, with little degeneration of system performance and little complexity in discovering the proper content to be recommended. The limited space in local memory of users’ terminal decides that only few content could be recommended and co-exist in users’ memory and the converging trend in accessing content mitigates such issue. Also, the fast decay in popularity of files make it possible for recommendation scheme to track the variation of hot content by fast updating the content in users’ terminal.

V. CONCLUSION

This paper presents the location based content recommendation for CASoRT system. Through analyzing the access log from Meepo site in Tsinghua campus network, the converging character in geography is revealed. The geographically independent content would cause the converging pattern of access in geography, as users’ imbalance distribution in geographic. Moreover, users tend to access to Meepo in fixed location. The users’ character of immobility makes recommending content to specific location relatively easy, for the component of user members in certain location would keep stable during a period of time. The analysis also reveals concentration of access in content dimension within a period of time. While most files scarcely visited, some hot files would be accessed in concentration and induce high amount of traffic in network. Therefore, with those properties of access, recommending those hot files to CASoRT system to broadcast throughout the region with high interest would alleviate the traffic load of wireless transmission in wireless network.

However, the model of users’ behavior is so complex that our work only reveals small perspective. Our location based content recommendation scheme could only discover the hot content after the access time accumulates to certain level. This scheme may miss the best time to recommend those hot files, as the fast change of access number of hot files in time domain. The method of forecasting the possible hot files should be studied.

ACKNOWLEDGEMENT

This work is supported by 973 National Basic Research Program of China (2012CB316000), Science Fund for Creative Research Groups of NSFC (61021001), National S&T Major Project (2013ZX03003004-002), 863 National High Technology Research and Development Program of China (2012AA011402), open research fund of state key laboratory of mobile communications (2012D02), and Tsinghua University Initiative Scientific Research Program (2010THZ03).

REFERENCES