UDC 621.391

ANALYSIS OF CONTENT REPLICATION MECHANISM IN RANDOM SEARCH IN PEER-TO-PEER NETWORKS

Volodymyr V. Popovskyy, Kateryna O. Popovska
Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

Background. Nowadays, file-sharing peer-to-peer (P2P) networks are becoming increasingly popular when providing video over the Internet. Successful reception of a given service is determined by how quickly and qualitatively the necessary content stored in the memory of the peers of this network is found. In structured systems a limited number of transitions are used to find the ordered file, however, for unstructured systems, where random search is used, the delay can be significant, especially when searching for rare files. At the same time, it is unstructured systems that can support complex requests with metainformation, which provides their priority development. There arises a question of how quickly the given file with a certain number of active peers can be obtained.

Objective. The aim of the paper is the analysis of the state of downloaded content on the nodes of unstructured P2P networks.

Methods. Synthesis of the mathematical model of video content replication taking into account the network load and performance analysis based on this model has been performed for the 3 most used algorithms: unified, proportional and “square root”. Numerical quantitative studies have been performed by means of calculation.

Results. A mathematical model that allows estimating the average number of attempts to replicate the ordered file taking into account the size of the network memory has been developed. The examples of the dependence of the average file search size for different numbers of active peers have been obtained.

Conclusions. The mathematical model of video content replication necessary for the successful search of the requested file is obtained taking into account the network memory size determined by the number of active peers. The obtained dependence of the average number of request attempts when searching for the desired files for the three replication algorithms allows concluding that the “square root” algorithm is expedient.

Keywords: peer-to-peer networks; mathematical models; files; performance; algorithm.

I. Introduction

Peer-to-Peer (P2P) file-sharing systems are overlay networks that work over the Internet [1,2].

A P2P network is a distributed system that contains interconnected nodes that are capable of self-organizing into the network topology for the purpose of sharing resources such as content, processor cycles, storage devices, and bandwidth adaptable to failures and a variable number of nodes while maintaining an acceptable level of connectivity and performance without the need for intermediaries or support of a global central server.

The range of such systems for content distribution starts from relatively simple applications for direct file sharing and extends to more complex ones that create distributed storage environments that provide security and efficient organization, indexing, searching, updating and retrieving data. Examples include the network PPLive, PPStream, Napster, Gnutella, Kazaa, Freenet, Groove, etc. [2,3,4].

Currently the P2P technology is considered to be a basis for a huge number of popular network services – from simple file sharing to voice and video communication, it is possible to perform distributed computing that allows users to use remote computers to perform complex data processing.

The dispersion of resources, which occurred due to the emergence of peer-to-peer networks, led to the virtual disappearance of costs associated with maintaining a giant centralized infrastructure.

The state of P2P networks is determined by how interacting peers self-organize into overlay networks and how they store or retransmit the requested files for each other. The main problem is to achieve effective resource search in a large-scale distributed storage network [1,2,3].

Currently, the popular P2P search systems can be classified as unstructured and structured depending on overlay structures [4,5]. Unstructured systems do not impose any structure on overlay networks. These systems are usually resistant to the dynamics of peers and support complex queries with metainformation. But they are not effective for finding rare unpopular files. Structured systems impose certain structures on overlay networks, which are commonly referred to as distributed hash tables (DHT). In a structured system, any file can be located in a small number of overlay transitions, which significantly reduces the search cost in comparison with unstructured systems. However, DHT only supports searching for one exact match keyword [6].

Depending on the level of decentralization of the system, we classify unstructured P2P networks into centralized, distributed, hybrid and analyze them in the conditions of the dynamics of functioning.

II. Main Part

Successful reception of the given service in P2P networks entirely depends on how quickly and qualitatively it is possible to find the necessary content stored in memory of peers of the given network. Content replication, therefore, plays a decisive role for successful searches. Obviously, having a large number of copies on the network will speed up the process of finding the right file. However, on the other hand, this will require more memory for storing content. Let us consider in more detail the mechanisms of replication in terms of a tradeoff between search efficiency and the cost of storing content.

In our unstructured P2P networks, we are essentially using a random search. In this case, in separate networks (Gnutella), peers requesting a file perform a random search and copy these files. In other networks (FastTrack) proactive replication is provided [6,7].

Suppose that in some P2P network there are m-files and n-nodes. Each i-file $i=(1 \le i \le m)$ is replicated at random peers $r_i = (1 \le r_i \le n)$. Suppose that the amount of memory for all peers is fixed and is the value of

$$R = \sum_{i=1}^{m} r_i \tag{1}$$

The probabilities of requests for each of the i-files are the equal and form a full group of events:

$$P = \sum_{i=1}^{m} P_i = 1 \tag{2}$$

The probability that the i-file is successfully located on the k th attempt:

$$P_i(k) = \frac{r_i}{n}\left(1 - \frac{r_i}{n}\right)^{k-1} \tag{3}$$

The average number of attempts to find the desired files is determined by the standard formula for estimating the average value

$$N = \sum_{i=1}^{m} P_i N_i = n \sum_{i=1}^{m} \frac{P_i}{r_i}, \tag{4}$$

where it is considered that the average number of requests for one file N_i is the value n/r_i.

Let us analyze the performance of three known replication algorithms.

1. Unified algorithm. In this algorithm all files are replicated regardless of their popularity, i.e.

$$r_i = \frac{R}{m}, \tag{5}$$

thus the average size of the search is

$$N_y = n \sum_{i=1}^{m} P_i \frac{m}{R} = \frac{nm}{R}. \tag{6}$$

2. Proportional algorithm differs in the fact that the replication creates more copies for more popular files, for which their probabilistic measure is considered.

$$r_i = RP_i, \tag{7}$$

Therefore, the average size of the search is

$$N_n = n \sum_{i=1}^{m} \frac{P_i}{RP_i} = \frac{nm}{R}, \tag{8}$$

that is, both considered algorithms give the same average search size.

3. Algorithm of square root replication. Using (4), we obtain:

$$r_i = \frac{R\sqrt{P_i}}{\sum_{j=1}^{m} \sqrt{P_i}} \tag{9}$$

Hence the average value is:

$$N_k = \frac{n}{R}\left(\sum_{i=1}^{m} \sqrt{P_i}\right)^2 \tag{10}$$

Fig. 1 presents the graph of the average search size depending on the number of files submitted m. It is also noted that the deviation of the total number of peers to the total volume is $(N/R) = const$. It is also obvious that $n > m$, since the number of submitted files cannot exceed the number of peers in the P2P network.

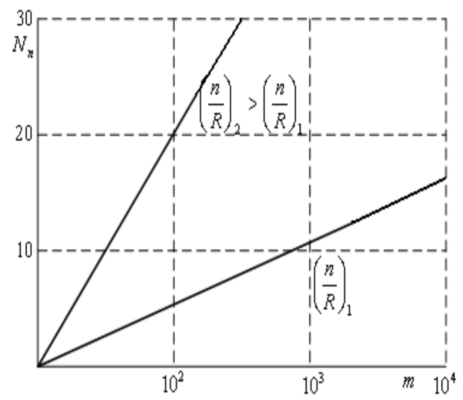

Fig. 1. The graph of the average search size depending on the number of files submitted

Fig. 2 shows the graphs of the average search size depending on the probability of requests P_i, where $P_i = const$ for all i.

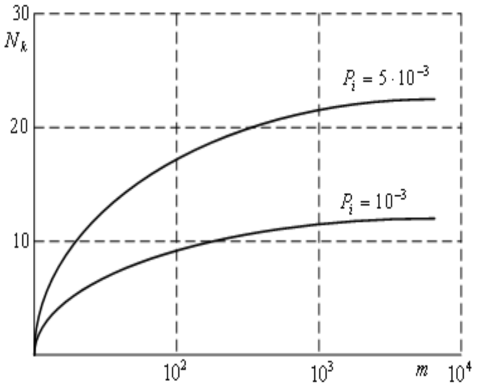

Fig. 2. The graph of the average search size depending on the number of requests for the "square root" replication algorithm

As follows from the graphs, the number of requests in the network is increasing with the increase in the number of files requested. However, for two algorithms: unified and proportional, this growth is linear, then for the "square root", where the demand for popular files is taken into account, this growth is not linear and more economical with increasing m .

III. Sammury

In unstructured P2P systems, unlike structured ones, addresses and content stored in a specific node are not interconnected. The search for the necessary information in these networks is carried out on the basis of two solutions:

- blind flooding, in which requests are sent to all connected nodes to check the requested files (Gnutella) in them;
- information search, in which peers are provided with additional information about the network topology and the location of resources (information awareness).

The advantage of blind flooding methods is that they do not require any service headers to maintain additional indices about the location of resources and are highly stable under dynamic loads. In contrast, information search methods allow nodes to support additional information about other nodes of the network, its topology and location of resources.

The essence of information methods is that nodes are supposed to collect and store topological and statistical information about their neighbors, which allows to optimize the topology, use adaptive algorithms for searching (Adaptive Probabilistic Search – APS). Each node saves the routing index, which stores detailed information about the contents of the files. Based on the routing indexes on the nodes, query routing tables are created by caching the file keywords. This allows networks to self-organize when searching for the right resources. Thus, with information search it is possible to dramatically increase the performance of finding the right resources.

IV. Conclusions

1. The mathematical model of video content replication necessary for the successful search of the requested file is obtained taking into account the network memory size determined by the number of active peers.
2. The obtained dependence of the average number of request attempts when searching for the desired files for the three replication algorithms allows concluding that the square root algorithm is expedient.
3. It is expedient to perform the necessary information on the basis of information search (information awareness, which is much more effective than the blind flooding).

References

1. Cisco Visual Netwroking Index; Forecast and Methodology, 2012-2017 Cisco Public, 2013.

2. Tarcoma S. Overlay Network Toward Information Networking – CRC Press, 2010.

3. Setton E., Girod B. Peer-to-Peer Video Streaming/Springer. – 2007. – 150 p.

4. Wehrle K., Gunes M., Gross J. Modeling and tools for Network simulation. Springer, Heidenberg, Dordrecht, London, Net York, 2010.

5. Дорт-Гольц А.А. Анализ функционирования наложенных сетей в сетях операторов. М. Электросвязь №3, 2014. Стр. 22-27.

6. Москалец Н.В., Поповская Е.О. Модели одноранговых пиринговых сетей для передачи потоков IPTV. Международный научно-исследовательский журнал, ISSN 2303-9868, №4, (23) 2014, Часть 2. Стр. 46-50.

7. Адами Амину, Гайдамака Ю.В. Аппроксимация вероятностных характеристик модели сети P2P. Л.2013. Инф.упр. сети. стр.4-10.

8. Popovskij V., Barkalov A., Titarenko L. Control and Adaptation in Telecommunication Systems, Springer-Verlag Berlin Heidelberg, 2011

Поповський В.В., Поповська К.О.

Аналіз механізму реплікації контенту під час випадкового пошуку в пірингових мережах

Проблематика. На сьогоднішній день файлообмінні пірингові P2P-мережі набувають все більшої популярності при наданні послуги відео через Інтернет. Успішне отримання заданої послуги визначається, тим наскільки швидко і якісно буде знайдений необхідний контент, що зберігається в пам'яті пірів даної мережі. Якщо в структурованих системах на пошук замовленого файлу витрачається обмежене число переходів, то для неструктурованих, де використовується випадковий пошук, затримка може виявитися чималою, особливо при пошуку рідкісних файлів. У той же час саме неструктуровані системи здатні підтримувати складні запити з метаінформацією, що забезпечує їм пріоритетний розвиток. Виникає питання: наскільки швидко вдається отримати заданий файл при певній кількості активних пірів.

Мета дослідження. Аналіз стану викачуваного контенту на вузлах неструктурованих P2P-мереж.

Методика дослідження. Проведено синтез математичної моделі реплікації відео контенту з урахуванням навантаження мережі і аналіз продуктивності на базі цієї моделі для 3х найбільш вживаних алгоритмів: уніфікованого, пропорційного і «квадратного кореня». Розрахунковим шляхом проведені чисельні кількісні дослідження.

Результати дослідження. Розроблено математичну модель, що дозволяє оцінювати середню кількість спроб реплікації замовленого файлу з урахуванням розміру пам'яті мережі. Отримано приклади залежності середнього розміру пошуку файлу при різних числах активних пірів.

Висновки. Математична модель реплікації відеовмісту, необхідна для успішного пошуку запитаного файлу, отримана з урахуванням розміру мережевий пам'яті, що визначається кількістю активних однорангових вузлів. Отримана залежність середнього числа спроб запиту під час пошуку потрібних файлів для трьох алгоритмів реплікації дозволяє зробити висновок, що алгоритм, який використовує механізм реплікації «квадратного кореня», є доцільним.

Ключові слова: пірингові мережі; математичні моделі; файли; продуктивність; алгоритм.

Поповский В.В., Поповская Е.О.

Анализ механизма репликации контента при случайном поиске в пиринговых сетях

Проблематика. На сегодняшний день файлообменные пиринговые P2P-сети приобретают все большую популярность при предоставлении услуги видео через Интернет. Успешное получение заданной услуги определяется, тем насколько быстро и качественно будет найден необходимый контент, хранящийся в памяти пиров данной сети. Если в структурированных системах на поиск заказанного файла затрачивается ограниченное число переходов, то для неструктурированных, где используется случайный поиск, задержка может оказаться значительной, особенно при поиске редких файлов. В тоже время именно неструктурированные системы способны поддерживать сложные запросы с метаинформацией, что обеспечивает им приоритетное развитие. Возникает вопрос: насколько быстро удается получить заданный файл при определенном количестве активных пиров.

Цель исследования. Анализ состояния скачиваемого контента на узлах неструктурированных P2P-сетей.

Методика исследования. Проведен синтез математической модели репликации видео контента с учетом нагрузки сети и анализ производительности на базе этой модели для 3х наиболее применяемых алгоритмов: унифицированного, пропорционального и «квадратного корня». Расчетным путем проведены численные количественные исследования.

Результаты исследования. Разработана математическая модель, позволяющая оценивать среднее количество попыток репликации заказанного файла с учетом размера памяти сети. Получены примеры зависимости среднего размера поиска файла при различных числах активных пиров.

Выводы. Математическая модель репликации видеосодержания, необходимая для успешного поиска запрошенного файла, получена с учетом размера сетевой памяти, определяемого количеством активных одноранговых узлов. Полученная зависимость среднего числа попыток запроса при поиске нужных файлов для трех алгоритмов репликации позволяет заключить, что алгоритм, использующий механизм репликации «квадратного корня», является целесообразным.

Ключевые слова: пиринговые сети; математические модели; файлы; производительность; алгоритм.