

ON THE APPROACHES OF INCREASING THE PRODUCTIVITY OF THE CORPORAL NET AND A CRITERION OF DATA UPDATING IN CASH-SERVERS

Amiraslan Aliyev¹, Aytan Huseynova²

Institute of Information Technology of ANAS, Baku, Azerbaijan

¹*amir@lan.ab.az*, ²*ayten@lan.ab.az*

The basis of any corporal net is local nets [1] in which new high speed technologies of data transmission are actively implemented. However the corporal net includes not only local computing net (LCN) but also many other divisions connected by dedicated channels or the general usage Internet net. Very often enterprises with territory distributed structures face the problem of consolidation of separated local nets which sometimes are significantly remote and differ in size and equipment.

For this purpose in the corporal nets the leased communication channels are used in which the speeds of data transmission are significantly lower as compared with the speeds of data transmission in local nets. Therefore the rational usage of the leased channels is of great significance for increasing the productivity of the corporal net.

The productivity of the corporal net can be increased by the following ways [2]:

-Acquiring the throughput capacity

The channel enhancement on the cost of acquiring additional throughput capacity solves the problem of the deficient productivity of the net for short-term period only as the traffic volume rises quickly and again begins to surpass the throughput capacity of the channels through which it is transmitted.

It can be explained very easily. In a nonoptimized net the need in the throughput band rises linearly towards to the user numbers (provided the applications are the same) [3].

-Managing the throughput capacity

The mechanisms of managing the throughput capacity allow better using the available throughput band due to its efficient distribution between various applications and users.

-Data compression

Data compression acts very efficiently for some file types (e.g., for the text files). However it is not of big use for transmission of "random" data such as preliminary compressed graphic files, archives, etc. When implemented in the real nets with package commutation the data compression means were found to be less efficient by both producers and users. In most cases the Internet or corporal net traffic compression increases the throughput capacity by up to 5-10%.

-Web Caching

The traditional technology of caching (Web Caching) allows allocating the static Web content closer to the ultimate user. However due to its nature this mechanism is applicable to the static Internet content only. And the main part of the traffic of the current corporal nets is on the contrary not related to the Internet and the intranets content is as a rule very dynamic one. The statistical research shows that caching the Web content increases the net productivity by about 30%.

- Content Delivery Networks (CDN)

CDN is a mode of the Internet access acceleration that showed itself positively in practice. They are overlaid nets via which the frequently requested Internet materials are relocated from the nucleus closer to the ultimate user. By transmitting the content to the net periphery the CDN service can complete subsequent requests for the same information faster as the number of intermediate routers of the global net reduces and delay diminishes, respectively. The content in the periphery is placed on the surrogate servers where the data are kept for subsequent extraction [4].

Of the intensive methods of the load optimization of the information transmission channels two approaches have really been developed: the information caching [5,6] and geometrical distribution of information within servers net (CDN – Content Delivery Network) [7].

Not only the CDN nets and caching mechanisms provide increasing the fastworking of access to the Web but diminish the load of the server – the data source. Other users when searching the same information will get in the cache nearest to them rather than to the source server. It will also reduce the consumption of the throughput capacity of the global net as the traffic moves from the ultimate user to the net periphery rather than to its nucleus. And the data caching in disks is much less expensive than upgrading the global net channels. The remote offices of the enterprise generally connect to their hosting centers through slower working connections. So, preliminary selection of content allows avoiding overload and minimizes the need for the corporal resources of the global net. Apart from that the CDN nets deliver the content of the source server into cache by one traffic flow and after that it divides into several ones for delivering into the desktop systems. In this connection the requirements for throughput capacity of the leased channel of the corporal net become still more moderate.

In the CDN nets processing the dynamic content is of importance. Accessing the data base in real timescale means decrease of the load of the servers with the original content. You can quickly create Web pages with access to the data base in real timescale and those needn't be remade every time when an updated or personalized content is requested. The result is that pages are generated with less excess, protection from the activity flashes is provided and demands level to the infrastructure reduces.

One of the urgent problems in the CDN nets is timely information updating on the surrogate servers. Most frequently the text information changes as compared with other kinds of information (audio, video). To define the updating criterion a measure based on the Levenstein metric is suggested.

The degree of the document change will be calculated by Levenstein distance (editing). Originally, the Levenstein distance was introduced for defining the distance between words. The Levenstein distance is equal to the minimal number of elementary editing operations needed for converting a line to another one. The set of elementary operations consists of the operations of substitution, insertion and deletion of one letter. To calculate the Levenstein distance every word should be presented as a set of symbols. The Levenstein distance between words $a = (a_1, \dots, a_N)$ and $b = (b_1, \dots, b_M)$ is calculated by the following recursive formula [8]:

$$L((a_1, \dots, a_N), (b_1, \dots, b_M)) = \begin{cases} N, & \text{if } M = 0 \\ M, & \text{if } N = 0 \\ \min \left\{ \begin{array}{l} L((a_1, \dots, a_{N-1}), (b_1, \dots, b_M)) + 1, \\ L((a_1, \dots, a_N), (b_1, \dots, b_{M-1})) + 1, \\ L((a_1, \dots, a_{N-1}), (b_1, \dots, b_{M-1})) + \bar{\delta}(a_N, b_M) \end{array} \right\}, & \text{in other cases} \end{cases} \quad (1)$$

In the formula (1) $\bar{\delta}(x, y)$ is negation of the Kronecker symbol:

$$\bar{\delta}(x, y) = \begin{cases} 0, & \text{if } x = y \\ 1, & \text{if } x \neq y \end{cases}$$

For every distance $L((a_1, \dots, a_i), (b_1, \dots, b_j))$, $i = 1, \dots, N$, $j = 1, \dots, M$ we use short-cut designation L_{ij} . Specifically, $L(a, b) = L_{NM}$. To calculate L_{NM} all distances L_{ij} , $i = 0, 1, \dots, N$, $j = 0, 1, \dots, M$, $(i, j) \neq (N, M)$ should be calculated. It is rather painstaking to calculate the Levenstein distance by the recursive formula. The complexity of such a calculation rises exponentially with word size. Instead of the recursive calculation the L_{ij} distance can be

calculated iteratively. In the iterative algorithm first the distances $L_{i0} = i$, $i = 1, \dots, N$ and $L_{0j} = j$, $j = 1, \dots, M$ are initialized. Then by the $L_{(i-1)j}$, $L_{i(j-1)}$ and $L_{(i-1)(j-1)}$ the distances L_{ij} , $i = 1, \dots, N$, $j = 1, \dots, M$ are sequentially calculated.

For calculation of the proximity of two text documents the Levenstein distance is not the adequate measure of proximity. The point is that rearrangement of single words in a sentence or sentences in a document is not perceived by human as a substantial mistake. Therefore when defining the proximity of two documents we will exclude the operation of substitution from the set of elementary operations. Let $L(D, D^*)$ be the minimal number of elementary operations of deletion and insertion of words which are necessary to convert a document in another one. Then the distance of proximity between the document D and copies D^* will be defined this way:

$$dist(D, D^*) = \frac{L(D, D^*)}{\lambda + \lambda^*}, \quad (2)$$

where λ and λ^* are lengths (in words) of the documents D and D^* , respectively.

From the definition we can see that $0 \leq dist(D, D^*) \leq 1$. The distance (2) between the identical documents is equal to zero and between different documents it is equal one.

References

1. “Current corporal nets” www.race.ru/page.asp?id=391
2. www.sovtel.ru/cgi-bin/products.pl?id=115
3. Joseph Yakubovich. “Optimisation of the net traffic”, Journal “Seti i Sistemy Svyazi”, N 10, 2001 (In Russian)
4. Doog Allen “Home nets of the content delivery” LAN, #01/2002 (In Russian).
www.osp.ru/lan/2002/01/
5. Ari Luotonen, Kevin Altis. World-Wide Web proxies. Proc. 1st Int. WWW Conf., Geneva, May 1994.
6. A. Chankhunthod, P. Danzig, C. Neerdaels, M. F. Schwartz, K. Worrell. A hierarchical Internet object cache. Proc. 1996 USENIX Annu. Tech. Conf., San Diego, Jan 1996, pp.153-163.
7. Алгулиев Р.М., Алыгулиев Р.М., Шарифов М.Г. Подход к оптимальному расположению серверов CDN в узлах глобальных сетей // Информационные технологии. 2006. № 11. С. 20-26.
8. Runkler T.A., Bezdek J.C., Web mining with relational clustering //International Journal of Approximate Reasoning. 2003. Vol. 32. №№2-3, pp. 217-236.