STATISTICAL MODELS FOR MINIMIZING THE NUMBER OF SEARCH QUERIES

I.V. Atlasov[1], *V.E. Bolnokin*[2], *O.Ja.Kravets*[3], *D.I. Mutin*[2], *G.N. Nurutdinov*[4]

[1] Moscow University of Ministry of Internal Affairs of Russian Federation named by V.Ja. Kikot
[2] Mechanical Engineering Research Institute of the Russian Academy of Sciences
[3] Voronezh state technical university; [4] Tambov state technical university
e-mail: csit@bk.ru
Russian Federation

Abstract: The problem of object monitoring is considered. It is necessary to constantly collect information that is quite rare over a large area and stored on various remote objects (servers). The fact that information is rare implies its special value. The current task is to minimize the total time for searching and processing information throughout the system. All remote servers are divided into groups of equal capacity. For each group, a server is allocated that will store up-to-date information about the availability of the necessary information. It is shown that if you first search for the availability of the necessary information on servers that control groups of servers, you can significantly reduce the search due to the fact that it is not necessary to search for the necessary information on all servers. You can only do this on servers that have traces of the necessary information and the group of servers that this server controls. The paper shows that using further grouping of servers, you can significantly reduce the search time for the necessary information.

Key words: statistical models, search queries, object monitoring, management.

1. INTRODUCTION

Historically, the heterogeneity of telecommunications systems, computer networks, network information resources, and the audience of their users complicates the objective monitoring and analysis of telecommunications architectures and resources. In this regard a fairly wide range of modern technical and technological solutions for monitoring and analysis should be used in the operation of telecommunications systems and computer networks (for example, radio monitoring [1], simulation of telecommunication systems [2], multi-levels heterogeneous routing [3], automatic monitoring & detection system for grey traffic [4], monitoring services for scalable heterogenous distributed systems [5].

Theoretical analysis of such systems is usually carried out on the basis of Queueing System theory [6], and various approaches are used – both a direct analytical description focused on a specific subject area [7], and a more General application for Markovian flow modeling of High-Reliability Telecommunications Systems [8], for network flows analysis [9], in large-scale (military) personnel systems [10], for transport of mixed-size sediment particles under unsteady flow conditions [11], for analysis of waiting time process of Markov type [12]. From the point of view of theoretical description, communication channels are usually interpreted as edges of the graph describing the system and are passive elements. Our article suggests a different approach.

Let's consider the task of monitoring objects. Let's assume that you need to constantly collect information that is quite rare over a large area and stored on various remote objects (servers). The fact that information is rare implies that it is of particular value. More precisely, the lower the p-probability of these events occurring, the greater the value of H = -ln(p) – the informative value of this event. Due to the importance of the necessary information, there is an urgent task – to increase the speed of collecting, processing the necessary information and protection. This task is solved for the entire server system.

The purpose of this article is to present an authors' proposal for solving the problem of minimization the total time spent searching and processing information throughout the system. In essence, this is a particular problem of the General task of information monitoring. This problem can significantly reduce the search time for the necessary information. The method of statistical modelling is used to solve the problem and the results show the effectiveness of the proposed approach.

The paper is organized as follows. The next section discusses the decomposition of main problem onto n-level problems. Section 3 presents new probabilistic method for solving the problem. The main results are presented in section 4. Section 5 with recommendations and example shows the effectiveness of the new method.

2. MATERIALS AND METHODS

The solution to this problem may be to divide all remote servers into groups of servers with k servers in each group. Next, you need to select the servers that will store up-to-date information about the availability of the information we need, more precisely, information that either the necessary information is missing, or there is at least one of the servers in the group. Dedicated servers will be called zero-level servers. If you first search for the availability of the necessary information on zero-level servers, you can significantly reduce the search, due to the fact that you do not have to search for the necessary information on all servers. You can only do this on those zero-level servers that have traces of the necessary information and the group of servers that this server controls. It is obvious that with a sufficiently small p, you can significantly reduce the time for searching and processing the necessary information. Let's take a closer look at one of the solutions to the original problem. So, we believe that we have n servers. Let Y_i be an integer random variable equal to the number of performed procedures for searching and processing information for the i group of servers. The total number of search queries is equal to $Y = \sum_{i=1}^{\frac{n}{k}} Y_i$. If the mathematical expectation is $M(Y) = \sum_{i=1}^{\frac{n}{k}} M(Y_i) = \frac{n}{k} M(Y_1) < n$, then we can assume that a certain amount of savings has been obtained. As a result of research on the extremum of the function $\varphi(k) = \frac{1}{k} M(Y_1)$ it is easy to prove that $\varphi(\frac{1}{\sqrt{p}}) = 2\sqrt{p}$.

It may happen that this time of searching for the necessary information is quite long. In this case, you can split all the remote servers of each k-th group into w groups of v servers in each. Next, you need to select the servers that will store up-to-date information about the availability of the necessary information. These servers will be called zero-level servers. In turn, the zero-level servers control the first-level servers, which store traces of the necessary information. Let's start searching for the necessary information on the zero-level servers. If there is data about the presence of the necessary information on the zero-level server, then we look for traces of the necessary information at a lower level, and only if we find these traces, we look for information directly on the servers where it is. You can build an analog of the random variable Y_i - an integer random variable θ_i - also equal to the number of search queries. Next, you need to investigate at least the function $M(\Theta) = \sum_{i=1}^{\frac{n}{k}} M(\Theta_i) = \frac{n}{k} M(\Theta_1)$. It is easy to prove that for function $\varphi(v, w) = \frac{1}{k} M(\Theta_1)$ the minimum is reached at the $v_0 = w_0 = \frac{1}{\sqrt[3]{p}}$ and is equal to $\varphi\left(\frac{1}{\sqrt[3]{p}}, \frac{1}{\sqrt[3]{p}}\right) = 3p^{\frac{2}{3}}$. It follows that at low p, the second method is more effective than the first. Similar questions were solved in [13, 14]. This work is also summarized in [15-17].

3. NEW PROBABILISTIC METHOD

The paper considers a method for solving the problem. A random variable equal to the number of search queries is generated and its mathematical expectation is minimized. Let's take some numbers v_1, v_2 and v_3 and denote $k=v_1v_2v_3$. Divide the entire set of n servers into k groups. Next, again divide the k servers into v_1 groups of v_2v_3 servers in each. Each group of v_2v_3 servers is divided into v_2 groups of v_3 servers in each. Next, we will perform the following actions.

1. A group of n servers, we split into groups according to v0 servers in each. Each group of v_0 servers is called a level 0 group. We also assume that the level 0 group contains $v_0= v_1v_2v_3$ elements. The number of zero-level groups is equal to n/v_0 (integer).

2. Each group of level 0 is divided into v_1 groups of level 1. Each level 1 group contains v_2v_3 servers.

3. Each group of level 1 is divided into v_2 groups of level 2. Each level 2 group contains v_3 servers.

4. Each element of the level 2 group will be called a level 3 server.

Let's look at how the system works. Assign each server a certain number (l,k,j,i), where l is the level 0 index, k is the level 1 index, j is the level 2 index, and i is the level 3 index. The total number of such servers is $v_0v_1v_2v_3$, where v_i is the number of servers of type i=0...3. Some of these servers store the necessary information, and it is not known which ones exactly. Moreover, we believe that this information may appear and also cease to be relevant for a fairly short period of time. Further information is provided in table 1.

Table 1. Method operation

Step (level)	Fixing	Main server	Task	There is the necessary information	There is no necessary information	Stored
1	l,k,j	(l,k,j,1) == (l,k,j)	send requests to servers with numbers (l,k,j,i) for fixed indexes l,k,j for all $1 \le i \le v_3$	{<l,k,j,i>=1}	{<l,k,j,i>=0}	{0, 1} (no or the necessary information is available on the group's servers $\{l,k,j,i\}_{i=1}^{v_3}$
2	l,k	(l,k,1) == (l,k)	send requests to servers with numbers (l,k,j) for fixed indexes l,k for all $1 \le j \le v_2$	{<l,k,j>=1}	{<l,k,j>=0}	{0, 1} (no or the necessary information is available on the group's servers $\{l,k,j\}_{j=1}^{v_2}$

3	l	(l,1)== (l)	send requests to servers with numbers (l,k) for fixed indexes l,k for all $1 \le k \le v_1$	{<l,k>=1}	{<l,k>=0}	{0, 1} (no or the necessary information is available on the group's servers $\{l,k\}_{k=1}^{v_1}$
4		(l) == (0)	send requests to servers with numbers (l) for fixed indexes l for all $1 \le l \le v_0$	{<l>=1}	{<l>=0}	{0, 1} (no or the necessary information is available on the group's servers $\{l\}_{l=1}^{v_0}$

Considered example. Let's assume that only the first x servers of the first level contain the following information in our dedicated servers:

$\{<11>=1\}\cup\ldots\cup\{<1x>=1\}\cup\{<1,x+1>=0\}\cup\ldots\cup\{<1,v_1>=0\}$.

Further, let's assume that for any $1 \le k \le x$, only the first x_k servers store 1 in the selected files, while the remaining files contain 0:

$\{<1,k,1>=\bullet\}\cup\ldots\cup\{<1,k,x_k>=\bullet\}\cup\{<1,k,x_k+1>=\bullet\}\cup\ldots\cup\{<1,k,v_2>=0\}$.

Further, for any $1 \le k \le x$ and x_k, let only the first x_k servers have the value 1 in the selected files:

$\{<1,k,x_k,1>=1\}\cup\ldots\cup\{<1,k,x_k,x_{kl}>=1\}\cup\{<1,k,x_k,x_{kl}+1>=0\}\cup\ldots\cup\{<1,k,x_k,v_3>=0\}$.

It does not matter which servers on the next third level contain the necessary information, since this does not affect the number of requests. The total number of requests to make is equal to

$$\zeta = v_0 + v_1 + xv_2 + v_3 \sum_{i=1}^{x} x_i \cdot$$

So, for $1 \le x \le v_1$, $1 \le x_i \le v_2$ we have

$$P(\zeta = v_0 + v_1 + xv_2(x_1 + x_2 + x_3)v_3) = C_{v_1}^{x}(1-p)^{(v_1-x)v_2v_3} \square \prod_{i=1}^{v_2} \left[C_{v_2}^{x_i}(1-p)^{(v_2-x_i)v_3}\overline{p} \right],$$

where $\overline{p} = 1-(1-p)^{V_3}$. Denote by the symbol v_2^x the set of all vectors $\overline{x} = (x_1, \ldots, x_x)$ of dimension x, each element x_j can take the values of a natural series within $1 \le x_j \le v_2$, $j=1\ldots x$.

In this paper, we construct an analog of the random variable Y_i - the random variable Ξ_i for all $1 \le x \le v_1$ and $\overline{x} = (x_1, \ldots, x_x) \in v_2^x$, which takes the form

$$\begin{cases} v_0, b & & (1-p)^{v_1 v_2 v_3} \\ \ldots & \ldots & \ldots \\ v_0 + v_1 + v_2 + v_3, & & \left[C_{v_1}^{1}(1-p)^{(v_1-1)v_2v_3}\right] \times \left[C_{v_2}^{1}(1-p)^{(v_2-1)v_3}\right] \times \\ 1 \le b \le v_3 & & \times(1-(1-p)^{v_3}) \\ \ldots & \ldots & \ldots \\ v_0 + v_1 + xv_2 + v_3\sum_{i=1}^{x} x_i & & \left[C_{v_1}^{x}(1-p)^{(v_1-x)v_2v_3}\right] \times \prod_{i=1}^{x}\left[C_{v_2}^{x_i}(1-p)^{(v_2-x_i)v_3}\right] \times \\ \sum_{i=1}^{x} x_i \le b \le v_3\sum_{i=1}^{x} x_i & & \times(1-(1-p)^{v_3})^{\sum_{i=1}^{x} x_i} \\ \ldots & \ldots & \ldots \\ v_0 + v_1 + v_1v_2 + v_1v_2v_3, & & (1-(1-p)^{v_3})^{v_1 v_2} \\ v_1v_2 \le b \le v_1v_2v_3 & & \end{cases}$$

where b is the number of search queries.

For the random variable $\Xi = \sum_{i=1}^{\frac{n}{V_1V_2V_3}} \Xi_i$ we have

$$M(\Xi) = \sum_{i=1}^{\frac{n}{v_1v_2v_3}} M(\Xi_i) = nH(v_1, v_2, v_3)$$

where

$$H(v_1, v_2, v_3) = \frac{1}{v_1v_2v_3}(1-p)^{v_1v_2v_3} + \frac{1}{v_1v_2v_3}\sum_{x=1}^{V_1}\left[C_{v_1}^{x}(1-p)^{(v_1-x)v_2v_3}\right] \times$$

$$\times \sum_{i=1}^{x}\left(\sum_{\overline{x}=(x_1,\ldots,x_x)\in v_2^x}\left(v_0 + v_1 + v_2 + v_3\sum_{j=1}^{x} x_j\right)\right) \times$$

$$\times \prod_{j=1}^{x} C_{v_2}^{x_j}(1-p)^{v_3\sum_{j=1}^{x}(v_2 - x_j)}\left(1-(1-p)^{v_3}\right)^{\sum_{j=1}^{x} x_j})$$

Let's go back to the tasks that were solved earlier. The rating is shown $\min_x \varphi(x) = 2p^{\frac{1}{2}}$. When splitting the interval [0, k] into smaller parts of the same length, the equation $\min_{x,y} \varphi(x, y) = 3p^{\frac{2}{3}}$ is obtained. Using the Stirling formula [3], we can prove that when $v_1 = v_2 = v_3 = \frac{1}{\sqrt[4]{p}}$ equality $H\left(\frac{1}{\sqrt[4]{p}}, \frac{1}{\sqrt[4]{p}}, \frac{1}{\sqrt[4]{p}}\right) = 4\sqrt[4]{p^3}$ is met.

By analogy, we can prove the statement.

Theorem. *If we consider s levels, then the estimate is valid*

$$M(\Xi_s) \approx n(s+1)p^{\frac{s}{s+1}}$$

provided that the power of each level is equal to $\frac{1}{\sqrt[s+1]{p}}$.

By taking the derivative and equating it to zero, we can show that the function $f(s) = (s+1)p^{\frac{s}{s+1}}$ reaches the minimum at the point s_0=[-ln(p)-1], and $\frac{1}{e^2} < p < (s_0+1)^{\frac{-s_0}{s_0+1}}$.

4. MAIN RESULTS

If we use the symbol Ξ_s to represent a random variable equal to the number of search queries at s levels and provided that the volume of each level is equal $\frac{1}{\sqrt[s+1]{p}}$, then as shown above $M(\Xi_s) \approx n(s+1)p^{\frac{s}{s+1}}$, it follows That the method can be used for $(s+1)p^{\frac{s}{s+1}} < 1$, or $p < (s+1)^{\frac{-s}{s}}$.

From the condition

$$\frac{\lim\limits_{p\to\infty} M(\Xi_{s+1})}{M(\Xi_s)} \approx \frac{\lim\limits_{p\to\infty}(s+2)p^{\frac{s+1}{s+2}}}{(s+1)p^{\frac{s}{s+1}}} = \frac{s+2}{s+1}\lim_{p\to\infty} p^{\frac{1}{(s+2)(s+1)}} = 0,$$

it should be noted that when the level increases, each subsequent method is more effective than the previous one.

5. RECOMMENDATIONS

1. Select s_0=-ln(p)-1.

2. Divide the entire set of servers into s_0 levels so that the volume of each level is equal to an integer part of the number $\frac{1}{\sqrt[s_0+1]{p}}$.

3. In this case, the mathematical expectation of the total number of searches is approximately equal to

$$M(\Xi_{s_0}) \approx n(s_0+1)p^{\frac{s_0}{s_0+1}} = n\min_s (s+1)p^{\frac{s}{s+1}}.$$

Considered example. Let p=0.001. It is obvious that for s_0=-ln(0.001)-1=5, the equality is fulfilled $(s_0+1)p^{\frac{s_0}{s_0+1}} \approx 0.01$. At the same p=0.001 and s=2,3,4 we get respectively $(s+1)p^{\frac{s}{s+1}} \approx 0.06$; 0.03; 0.02. This shows the effectiveness of the new method at s_0=5 in six, three and two times, respectively.

6. CONCLUSION

The problem of reducing the number of search queries has been solved. To compare the results obtained, the history of solving the problem was considered. It is shown that if p<0.5 is the probability of finding the necessary information on a given server and there are n servers, then you will have to make about $2np^{\frac{1}{2}}$ queries to find the necessary information. It is also shown that by creating another level, you can reduce requests to $3np^{\frac{2}{3}}$. It is proved that when choosing the number of levels s=3, it is possible to achieve that the mathematical expectation of queries becomes equal $4np^{\frac{3}{4}}$. Moreover, it can be proved by analogy that when s levels are selected, the number of queries is equal $(s+1)np^{\frac{s}{s+1}}$.

REFERENCES

[1] Spajic V., Kozic N., Pokrajac I., Okiljevic P. Radio monitoring of telecommunication systems with TDMA multiple access technique. *Proc. of Telecommunications Forum (TELFOR)*, 2012.

[2] Lencse G., Derka I., Muka L. Towards the efficient simulation of telecommunication systems in heterogeneous distributed execution environments. in *Proc. of Telecommunications and Signal Processing (TSP), 2013 36th International Conference on.*

[3] Tyagi S., Tanwar S., Gupta S.K. et al. A lifetime extended multi-levels heterogeneous routing protocol for wireless sensor networks. *Telecommunication Systems*, Volume 59, Issue 1, 2015, pp. 43-62.

[4] Khan M.A., Imtiaz S.Y., Shakir M. Automatic Monitoring & Detection System (AMDS) for Grey Traffic. *Proc. of the World Congress on Engineering and Computer Science*, Vol II, 2015, San Francisco, USA, pp. 696-700.

[5] Subramanian et al. Gossip Enabled monitoring services for scalable heterogenous distributed systems. *Cluster computing*, 2006.

[6] Romansky R. An Application of Procedure for Analytical Evaluation of Computer Structures. *International Journal on Information Technologies and Security,* No 2 (Vol. 5), 2013.

[7] Norris J.R. Continuous-time Markov chains I. *Markov Chains*, 1997, p. 60.

[8] Kitchin J. Approximate Markov Modeling of High-Reliability Telecommunications Systems. *IEEE Journal on Selected Areas in Communications*, Volume 4, Issue 7, 2006, pp. 1133-1137.

[9] Morley C.D., Thornes J.B. A Markov Decision Model for Network Flows. *Geographical Analysis*, Volume 4, Issue 2, 2010.

[10] Merck J.W., Hall K. A Markovian flow model: the analysis of movement in large-scale (military) personnel systems. *Rand Corp*., 1971. R-514-PR.

[11] Kuai K., Tsai C. Discrete-Time Markov Chain Model for Transport of Mixed-Size Sediment Particles under Unsteady FlowConditions. *Journal of Hydrologic Engineering*, Volume 21, Issue 11, 2016.

[12] Clarke A.B. A Waiting Time Process of Markov Type. *Ann. Math. Statist*., vol. 27, 1957. - pp. 452-459.

[13] Lagutin M.B. Visual mathematical statistics. Moscow: BINOM, 2009.

[14 Atlasov I.V., Kravets O.Ja., Sekerin V.D., Gorokhova A.E., Gasanbekov S.K. Creating a model of resource saving for diagnosing various deviations. *Compusoft, An International Journal of Advanced Computer Technology*, 2019, 8(10), pp. 3440-3443.

[15] Hakrama I., Frashiri N. Agent-Based Modelling and Simulation of an Artificial Economy with Repast. *International Journal on Information Technologies & Security*, N. 2 (vol. 10), 2018.

[16] Hassani M.M., Berang R. An Analytical model to calculate blocking probability of secondary user in cognitive radio sensor networks. *International Journal on Information Technologies & Security*, N. 2 (vol. 10), 2018

[17] Atlasov I.V., Dubinina N.M. Building a model that allows saving resources necessary for diagnosing self-propagating organisms in a biologically active environment. *Control systems and information technologies*. 2017. Vol. 69. N. 3. P. 49-53.

Information about the authors:

Igor Viktorovich Atlasov – professor of Moscow University of Ministry of Internal Affairs of Russian Federation named by V.Ja. Kikot', areas of scientific research – system analysis, optimization, simulation of complex objects

Vitaly Evgenievich Bolnokin – professor of Mechanical Engineering Research Institute of the Russian Academy of Sciences, areas of scientific research – system analysis, neural networks, modeling

Oleg Jacovlevich Kravets – professor of Voronezh state technical university, areas of scientific research – system analysis, optimization, simulation of complex objects, networks

Denis Igorevich Mutin – professor of Mechanical Engineering Research Institute of the Russian Academy of Sciences, areas of scientific research – system analysis, neural networks, modeling

Gennady Nurislamovich Nurutdinov – associate professor of Tambov state technical university, areas of scientific research – system analysis, cyber security, mobile data transmission systems

Manuscript received on 29 June 2020