УДК 004.051

ВИРІШЕННЯ ПРОБЛЕМИ ДОСТУПНОСТІ ДО МЕРЕЖІ ІНТЕРНЕТ ШЛЯХОМ ДИНАМІЧНОГО БАЛАНСУВАННЯ ПРОПУСКНОЇ ЗДАТНОСТІ КАНАЛУ WEB-ТРАФІКУ

Юрій Яремчук, Дмитро Кец, Тетяна Жевега, Кирило Безпалий
Вінницький національний технічний університет

Анотація: **Розглянуто можливість вирішення проблеми доступності до глобальної мережі шляхом динамічного балансування пропускної здатності каналу. Запропоновано метод прогнозування навантаження в мережі, що базується на попередньому аналізі фактичних даних про використання мережі користувачами і розрахунком лінійного тренду та показників сезонності. Практичне застосування запропонованого методу показало високу точність прогнозу та можливість його застосування в реальних великих за розмірами мережах.**
Summary: **The paper considers the possibility of solving problems of access to global networks by balancing dynamic bandwidth. The method of forecasting the load on the network, based on preliminary analysis of evidence on the use of network users and the calculation of the linear trend and seasonal indices. The paper presented the practical application of this method, which showed high accuracy of prediction and possibility of its application to real large-size networks.**
Ключові слова: **Проблема доступності, методи прогнозування, динамічне балансування.**

І Вступ

З розвитком інформаційних технологій все більш широкого застосування набувають локальні мережі з наявним підключенням до глобальної мережі Інтернет. При цьому, в багатьох випадках, пропускна здатність каналу доступу до Інтернет є обмеженою як економічними, так і фізичними ресурсами підприємства. Зі зростанням кількості споживаної інформації в комп'ютерних мережах виникають пікові навантаження. При виникненні таких ситуацій ефективна пропускна здатність каналу прямує до нуля, що призводить до неможливості виконання своїх функцій мережею і відповідно унеможливлює роботу користувачів. Якщо детально проаналізувати проблему, то її причиною є відсутність механізмів автоматичного розподілу каналу з урахуванням зростаючих чи спадаючих потреб користувачів. Прикладом такої проблеми є користувачі з безлімітними квотами доступу до мережі Інтернет, які під час роботи в Інтернеті можуть займати всю пропускну здатність мережі і, як наслідок, в інших користувачів пропускна здатність каналу буде наближатися до нуля.

При аналізі методів прогнозування, побудованих на нейронних мережах [1], було виявлено їх основні недоліки, зокрема такі як індивідуальний підхід до складності мережі, пропорційний до її розміру, неможливість розробки універсальних правил для всіх типів мереж, досить довгий час, необхідний для навчання системи.

Використання статистичних методів прогнозування, таких як метод екстраполяції, апроксимації, регресії [2] потребують для аналізу та прогнозування досить багато даних [3], що, в свою чергу, є досить проблемним за відсутності даних за останні як мінімум два роки роботи, а це стає неможливим з урахуванням доцільності та об'ємів зберігання на фізичних носіях. Також значним недоліком web-трафіку є відсутність якості самоподібності [4], що може призвести до неточних або взагалі неправдоподібних прогнозів.

Існуючі на сьогодні програмно-апаратні комплекси, зокрема такі як Kerio, Microsoft TMG, Traffic Inspector можуть розподіляти наявний канал доступу до глобальної мережі Інтернет та відслідковувати, з якою метою і наскільки продуктивно використовується виділена пропускна здатність каналу, а також володіють можливістю відстеження та класифікації ресурсів, на які було використано трафік. Деякі з цих комплексів дозволяють статично розподілити навантаження на мережу з додатковим індивідуальним розширенням квот для кожного користувача. Проте жоден з них не дозволяє виконувати динамічне розподілення пропускної здатності каналу, бо це вимагає додаткового та детального аналізу трафіку, що вже було використано, та на основі отриманих результатів виконувати перерозподіл пропускної здатності наявного каналу. Аналіз інформації при цьому вимагає наявності великої кількості людських ресурсів та чимало часу. На основі отриманих результатів аналізу відповідальна особа чи системний адміністратор може вносити відповідні статичні зміни в мережеве обладнання, але навіть такий розподіл може призвести до перевантаження каналу.

Отже актуальною стає розробка методу забезпечення доступності до глобальної мережі Інтернет, який би вирішував вказані проблеми.

ІІ Метод забезпечення доступності до Інтернет на основі прогнозування навантаженості мережі

Пропонується метод забезпечення доступності до глобальної мережі Інтернет користувачами локальних мереж, який базується на прогнозуванні використовуваного трафіку як всією мережею так і окремими її сегментами, підмережами. Прогнозування базується на використанні попередньо зібраних даних про використання пропускної здатності каналу користувачами мережі шляхом статистичного аналізу попередньо зібраної інформації.

Однією з умов при прогнозуванні є розрахунок коефіцієнту кореляції між вибраними даними за досліджуваний період для знаходження статистичної залежності та оцінки похибки прогнозованих даних. Також рекомендовано з метою підвищення точності прогнозування виконувати відбір активних періодів дослідження, що базується на порівнянні даних вибірки та середньоарифметичним показником цієї вибірки. Безпосередньо процес прогнозування використовує показник тренду лінійності, що дозволяє визначити зміни динамічного ряду, індекс сезонності та коефіцієнт сезонності. Прогнозування базується на аналізі часових рядів роботи мережі за певний період, зокрема день тижня, година доби. Прогнозування на основі аналізу часових рядів передбачає, що зміни, які відбувалися в обсягах використання мережевого трафіку, можна використовувати для визначення цього показника в наступні періоди часу. Тимчасові ряди, зазвичай, служать для розрахунку трьох різних типів змін в показниках: трендових, сезонних (день тижня, година доби) і циклічних. Зазвичай, ці методи використовуються для прогнозування обсягів продаж в економіці та маркетинговій діяльності. Найбільш точні результати ці методи дають при прогнозуванні на короткі терміни та при наявності поняття сезонності. Слід зазначити, що показник сезонності в комп'ютерних мережах можна відслідковувати також по тижнях, по місяцях, по роках, але це вимагає наявності інформації про використання пропускної здатності каналу за досить великі проміжки часу, що є складним у реалізації.

Першим етапом є збір інформації з локальної мережі про запити до глобальної мережі Інтернет для подальшого прогнозування. Досить важливим питанням є довжина періоду, на який потрібно спрогнозувати потреби використання мережі. Для прогнозування на досить великий час потрібно мати дані, які як мінімум в 5 разів довші за прогнозований період, крім цього слід враховувати сторонні фактори, зокрема події у світі різного походження, вірусні активності, розвиток соціальних мереж, тощо, які можуть вплинути на тенденцію використання каналу доступу до мережі Інтернет. Слід відзначити, що система є досить залежною від людського фактора, що унеможливлює точний прогноз на досить великий період. Також слід зауважити, що основним призначенням таких систем є розподіл каналу доступу до мережі Інтернет між усіма користувачами для забезпечення роботи в пікові навантаження на канал доступу до мережі Інтернет.

Враховуючі всі вищеописані зауваження можна зробити висновок про те, що термін, на який слід робити прогноз, має бути мінімальним (день або тиждень). Під час прогнозування слід відкидати дні з досить малою активністю, бо вони не несуть інформованості про перевантаженість каналу і можуть вносити похибку для подальших прогнозів. Також, враховуючи розміри мереж, слід аналізувати дані в порядку поділу – від великих мереж до малих підмереж. Спочатку загальний аналіз всієї мережі, після чого більш детальний аналіз підмереж.

Збір інформації про використання пропускної здатності каналу з метою подальшого аналізу можна здійснювати різними способами. Найбільш поширеними серед них є збір логів Squide на мережевому шлюзі та логів Apache програмного мережевого екрану Kerio. Один із найпоширеніших способів збору інформації про активність в мережі є використання програмного забезпечення, яке дозволяє з мережевого шлюзу збирати дані про передану інформацію. З усіх даних, які може надавати мережевий шлюз, потрібно збирати такі: дата та час виконання запиту (з точністю до секунд), ідентифікатор мережевого рівня (IP – адреса), час, витрачений на виконання транзакції користувачем під час кожного запиту, кількість інформації отриманої користувачем (у байтах).

За рахунок цих даних існує можливість відслідковування активності роботи користувача протягом робочого дня/тижня/місяця, підрахувати кількість запитів та визначити об'єми завантаженого контенту. Однак, у мережах підприємств у зв'язку з великими обсягами даних у звітній інформації виникає проблема в можливості доступу до них та їх обробки з метою подальшого аналізу. Щоб вирішити її було розроблено комп'ютерну програму «SquidParser», яка дозволяє виконувати перенесення звітів з використання доступу до глобальної мережі Інтернет по протоколу HTTP до бази даних MS SQL Server. Алгоритм роботи програми виконаний у багатопоточному режимі, що значно підвищує швидкість роботи програмного модуля та збільшує швидкість перенесення інформації за рахунок використання багатоядерних процесорів.

Для подальшого прогнозування отриманні дані потрібно згрупувати у певній послідовності. А саме: на активність кожного клієнта відповідно до діб, годин, та підсумувати об'єм трафіку за кожну годину. Враховуючи той факт, що інформація для аналізу буде використовуватись в розрізі годин, а зібрані дані

записують після закінчення запиту користувача, можуть виникнути ситуації, при яких користувачі починають запит в одну годину, а відповідь отримують в іншу. Тому слід перерозподіляти завантажену інформацію у відсотковому співвідношенні по відповідним годинам. Для правильного перерозподілу слід використовувати такі показники, як час, витрачений на виконання транзакції, та час її закінчення.

Наступним етапом є формування таблиць з даними таким чином, щоб можна було з легкістю опрацьовувати інформацію згідно з методами аналізу, а саме: дата (рррр/мм/дд), IP-адреса, кількість інформації отриманої користувачем, кількість запитів.

Перед початком виконання прогнозування необхідно визначити статистичну залежність таких даних, як подібність навантаження мережі протягом робочого тижня по дням, а також у погодинному розрізі доби, через підрахунок їх коефіцієнта кореляції за формулою:

$$r_{k,L} = \frac{\sum_{i=1}^{n}(x_{i,k} - \overline{x}_k)\cdot(x_{i,L} - \overline{x}_L)}{\sqrt{\sum_{i=1}^{n}(x_{i,k} - \overline{x}_k)^2 \cdot \sum_{i=1}^{n}(x_{i,L} - \overline{x}_L)^2}}, \quad (1)$$

де n – кількість спостережень; $x_{i,k}$ спостереження i-ї змінної k; $\overline{x}_k$ – середнє значення k-ї змінної; k та L – чинники, між якими розраховується показник тісноти лінійного статистичного взаємозв'язку.

Коефіцієнт кореляції показує тісноту лінійного статистичного взаємозв'язку між двома співзалежними ознаками (величинами). У випадку аналізу мережі ці величини – це кількість трафіку за певні проміжки часу. Значення коефіцієнта кореляції вказує на величину похибки прогнозування [5].

Для підвищення точності результатів прогнозування потрібно провести аналіз роботи мережі на активність. Визначення активності проводиться порівнянням фактичних значень використанням мережі з їх середньоарифметичним за досліджуваний період часу. Всі значення, що менші за середнє арифметичне, не слід враховувати при подальшому аналізі та прогнозуванні. Середнє арифметичне знаходиться за формулою:

$$\overline{P} = \frac{\sum p}{n}, \quad (2)$$

де p - значення рівнів ряду; n - кількість рівнів ряду.

Малоактивні періоди рекомендується не враховувати через негативний вплив у вигляді зменшення точності прогнозу на майбутнє. Якщо є необхідність прогнозування малоактивних періодів, то це слід робити за тим же методом, але окремим блоком.

Дослідження основної тенденції розвитку є емпіричним прийомом попереднього аналізу. Для того, щоб дати кількісну модель змін динамічного ряду, використовується метод аналітичного вирівнювання, який в собі передбачає розрахунок тренду та індексу сезонності.

З метою оцінки тренду можуть використовуватися такі функції [6]:

- при рівномірному розвитку - лінійна функція, знаходиться за формулою:

$$Y_t = b_0 + b_1 t, \quad (3)$$

де Y – послідовність значень, що аналізуються; t – номер періоду часового ряду; $b_0, b_1 t$ – невідомі, які необхідно знайти;

- при зростанні з прискоренням – парабола другого порядку $Y_t = b_0 + b_1 t + b_2 t^2$ та кубічна парабола $Y_t = b_0 + b_1 t + b_2 t^2 + b_3 t^3$;

- при постійних темпах зростання – показова функція $Y_t = b_0 e^{b_1 t}$;

- при зниженні з уповільненням – гіперболічна функція $Y_t = b_0 + b_1 \frac{x_1}{t}$.

Параметри $b_0, b_1, ..., b_n$ знаходяться за методом найменших квадратів. Сутність методу полягає в тому, щоб знайти такі параметри b, при яких сума квадратів залишків буде мінімальною.

Для згладжування часових рядів недоцільно використовувати функції, що містять велику кількість параметрів, оскільки отримані таким чином рівняння тренду (особливо при малому числі спостережень) відображатимуть випадкові коливання, а не основну тенденцію розвитку явища.

Для кожного з елементів періоду розраховується коефіцієнт сезонності для періоду (день, місяць) та загальний індекс сезонності для всього досліджуваного періоду. У найпростішій формі індекс сезонності розраховується як відношення середнього рівня за відповідний місяць до загального середнього значення показника за рік згідно з формулою:

$$l_c = \frac{\overline{P_l}}{\overline{P_0}} \cdot 100, \tag{4}$$

де $\overline{P_l}$ – середнє з фактичних рівнів одноіменних проміжків часу; $\overline{P_0}$ – постійна середня за період, що вивчається.

Значення постійної середньої знаходиться за формулою:

$$\overline{P_0} = \frac{\sum \overline{P_t}}{n}. \tag{5}$$

Коефіцієнт сезонності розраховується за формулою:

$$K_c = \frac{N_i}{S_z}, \tag{6}$$

де N_i – фактична сума за певний період (місяць, день); S_z – середньомісячне значення за рік (середньоденне за тиждень).

Відповідно до отриманими даних відбувається розрахунок прогнозу на майбутній період за формулою:

$$Y_{t+v} = (u_t + b_t \cdot v) \cdot I_{t-s+v}, \tag{7}$$

де u_t – декомпозиційний ряд; v – період упередження (прогнозування); Y_{t+v} – скоригований на сезонність рівень прогнозу; b_t – характеристика лінійного тренду; I_{t-s+v} – скорегований індекс сезонності.

Декомпозиційний ряд знаходиться за формулою:

$$u_t = \frac{Y_t}{I_{t-s}}, \tag{8}$$

Головною проблемою при розподілі пропускної здатності каналу є пропорційність, при якій користувачі за необхідності більшої кількості запитів до глобальної мережі мали б можливість її безперешкодного отримання, але і не допускати виділення певного розміру пропускної здатності каналу користувачам, які не мають в цьому необхідності. Для вирішення цієї проблеми необхідно проводити аналіз завантаженості загальної пропускної здатності каналу всією мережею, кожною підмережею, а, у разі необхідності, і кожним комп'ютером у мережі.

III Застосування методу

Далі розглянемо застосування даного методу на прикладі конкретної мережі підприємства.

Дані, що використовуються при аналізі та прогнозуванні, були зібрані за допомогою програмного забезпечення Squid. Формування таблиць з метою подальшого аналізу відбувається з використанням згаданої вище розробленої комп'ютерної програми «SquidParser».

Першим кроком є визначення активності. Для цього було зроблену вибірку даних за три тижні. Результати проведеного дослідження наведені в табл. 1.

Таблиця 1 – Перевірка аналізованого періоду на активність

День тижня	Кількість трафіку (байт)	Активність	Кількість трафіку (байт)	Активність	Кількість трафіку (байт)	Активність
Понеділок	56306007373	А	58999273206	А	76794288698	А
Вівторок	65603040376	А	49986227598	А	71678852998	А
Середа	49820815192	А	53842491564	А	76430031645	А
Четвер	61626574054	А	58616250479	А	71760274395	А
П'ятниця	61685525489	А	54581206778	А	53771745667	НА
Субота	43728651550	НА	31129414890	НА	55820052500	НА
Неділя	3123679614	НА	2375400638	НА	2375400638	НА
Середнє	48842041950		44218609308		58436782383	

Примітка: А – активний; НА – не активний.

Згідно з отриманими даними є можливість побачити, над якими даними буде проводитись подальший аналіз з метою прогнозування.

Проведемо аналіз та виконаємо прогнозування на тиждень на прикладі роботи тієї ж мережі. Дані взято терміном за $N = 55$ активних днів. Розраховано коефіцієнт кореляції згідно з формулою (1), значення якого свідчать про можливість прогнозування з точністю до 10% [5]. Дані, на основі яких виконуватиметься прогнозування, наведені в табл. 2.

Таблиця 2 – Вибірка даних для прогнозування

Дата	Кількість отриманої інформації (Байт)
2012-01-16	63356622943
2012-01-17	49906874718
2012-01-18	50160559042
2012-01-19	67216553819
2012-01-20	47669135023
…	…
2012-03-23	73514031331
2012-03-26	78180507715
2012-03-27	65134271867
2012-03-28	70568321720
2012-03-29	65782791564
2012-03-30	69829821228

На рис. 1 наведено графічне відображення використання пропускної здатності каналу мережі за кількістю отриманої інформації за досліджуваний період (55 днів).

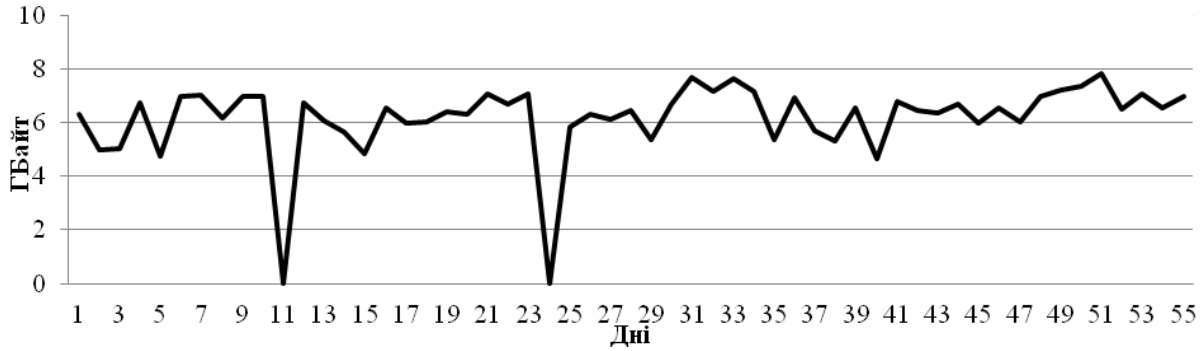

Рисунок 1 – Графічне відображення завантаженості мережі

При аналізі вибраних даних на рис. 1 спостерігається рівномірний лінійний розподіл трафіку. Це дозволяє стверджувати, що для розрахунку тренду необхідним є використання лінійної функції розподілу згідно з (3), що дозволить отримати максимально точні значення прогнозу.

Результатом розрахунку тренду є дані, що наведено в табл. 3.

Таблиця 3 – Розрахунок тренду та відхилення фактичних даних

Тренд	Відхилення фактичних даних від тренду
55 528 813 650	1,140968423
55 760 929 342	0,895015117
56 689 392 109	1,231954527
…	…
67 134 598 236	1,164533784
67 366 713 928	0,966861348
67 598 829 619	1,043928158
67 830 945 311	0,969805024
68 063 061 003	1,025957696

Слід зазначити, що значення тренду є взаємозалежною величиною з фактичними даними використання пропускної здатності каналу. Це пояснює досить великі значення тренду, бо кількість трафіку взято в кілобайтах.

Обраховуємо коефіцієнти сезонності для кожного дня за формулою (6) та загальний індекс сезонності для досліджуваного нами періоду за формулою (5). Їх значення наведено в табл. 4.

Таблиця 4 – Отримані результати коефіцієнта сезонності

Назва дня	№ дня тижня	Коефіцієнт сезонності	Індекс сезонності
Понеділок	1	1,02	61795937327
Вівторок	2	1,02	
Середа	3	1,03	
Четвер	4	0,96	
П’ятниця	5	0,97	

На основі проведених вище розрахунків виконуємо прогнозування на майбутній тиждень за формулою (7). Результати прогнозування наведені в табл. 5.

Таблиця 5 – Отриманий прогноз

Назва дня	№ дня тижня	Тренд	Отриманий прогноз
Понеділок	1	68 295 176 695	69 408 143 588
Вівторок	2	68 527 292 386	69 980 330 603
Середа	3	68 759 408 078	70 993 108 608
Четвер	4	68 991 523 770	66 331 542 729
П’ятниця	5	69 223 639 461	67 047 887 409

Порівняння фактичних значень використаного трафіку з отриманим результатом прогнозування відображено на рис. 2.

Рисунок 2 – Графічне відображення прогнозованих даних

Використовуючи описаний вище метод можна виконувати прогнозування навантаження в мережі та кожній підмережі за добами, але умовою є проведення прогнозування для активних годин дня (робочий час).

Проводимо аналіз завантаженості пропускної здатності каналу на прикладі досліджуваної мережі. Варто зауважити, що пропускна здатність каналу становить 50 Мбіт/с, що дорівнює можливості отримання близько 22500 Мбайт даних за годину на всю мережу. Провівши підсумовуючі обрахунки отримуємо дані наведені в табл. 6, які відображають завантаженість мережі з інтервалом у одну годину в розрізі однієї досліджуваної доби.

Таблиця 6 – Кількість отриманих даних мережею в кожну годину дня

Години	%	Мбайт	Години	%	Мбайт
0	≈ 0,00	0,065	12	19,75	4443,387
1	≈ 0,00	0,047	13	33,33	7499,022
2	≈ 0,00	0,098	14	27,56	6201,945
3	≈ 0,00	0,041	15	20,34	4577,134

4	≈ 0,00	0,039	16	8,27	1861,044
5	0,17	37,903	17	2,98	669,975
6	6,12	1377,333	18	2,61	587,927
7	19,07	4290,070	19	0,60	135,446
8	23,45	5275,464	20	≈ 0,00	1,025
9	23,40	5264,192	21	≈ 0,00	0,029
10	21,91	4929,484	22	≈ 0,00	0,817
11	29,09	6545,068	23	≈ 0,00	0,038

Графічне відображення ситуації, що склалась, показано на рис. 3.

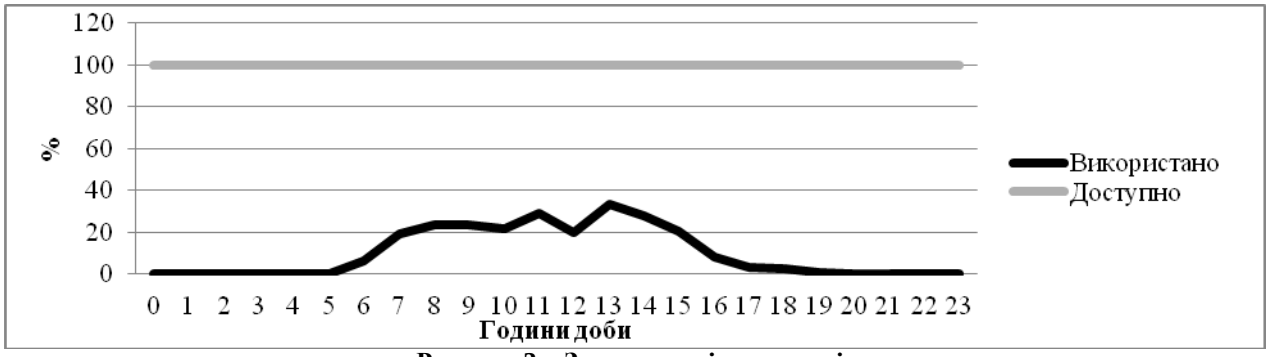

Рисунок 3 – Завантаженість мережі

Згідно з отриманими даними видно, що за день максимальний відсоток використання становить близько 34%, що свідчить про недоцільність отримання такої пропускної здатності каналу. У зв'язку з тим, що пропускна здатність каналу в мережі статично розподілена між підмережами, то при такій навантаженості можна говорити про перерозподіл пропускної здатності каналу з метою підвищення її для кожної підмережі.

Для повної впевненості в доцільності використання такої пропускної здатності каналу варто після переналаштування мережевого обладнання проводити подальший аналіз навантаженості. У випадку, якщо ситуація не змінилась, варто задуматись про необхідність такого каналу для підприємства.

IV Висновок

Проведено аналіз проблеми доступності до глобальної мережі Інтернет шляхом динамічного балансування пропускної здатності каналу між користувачами локальної мережі. Для вирішення цієї проблеми запропоновано метод прогнозування навантаження в мережі, що використовує аналіз попередньо зібраної інформації про використання доступу до мережі Інтернет користувачами за попередні періоди часу та виконує прогнозування навантаження на мережу на майбутні періоди.

Практично було показано можливість використання даного методу при роботі великих мереж з забезпеченням досить високої точності прогнозування. Запропонований метод дозволяє також виконувати аналіз завантаженості всієї мережі з метою виявлення «вільної» пропускної здатності каналу або високої завантаженості мережі, що варто враховувати при динамічному балансуванні пропускної здатності каналу між користувачами на основі прогнозованих даних.

Література: ***1.*** *Крисилов В. А. Чумичкин К. В. Кондратюк А. В. Представление исходных данных в задачах нейросетевого прогнозирования // Конференция «Нейроинформатика 2003». – М.:НАУЧНАЯ СЕССИЯ МИФИ, 2003. - с184-191.* ***2.*** *Семенов С. Г. «Анализ методом прогнозирования в телекоммуникационных сетях автоматизированных систем управления». Системи управління, навігації та зв'язку. – 2008, №2(6) – с. 134-137.* ***3.*** *Каграманзаде А. Г. Прогнозирование и проектирование телекоммуникационных сетей: Монография. - Баку: Бакинский Университет, 1998. - 242 с.* ***4.*** *Петров В. В. Структура телетрафика и алгоритм обеспечения качества обслуживания при влиянии эффекта самоподобия. Автореферат диссертации. Москва, 2005. - 20 с.* ***5.*** *Петрук В. А., Кашканова Г. Г. Ймовірнісно-статистичні моделі та статистична оцінка рішень. Навчальний посібник – Вінниця: УНІВЕРСУМ-Вінниця, 2006 – 131 с.* ***6.*** *Астафурова И. С. Статистика. Учебно-методическое пособие. / Под редакцией Масленникова С. Г. – М: ВолГУ – 1998 г. – 24 с.*