A Novel Wavelet-Based Approach to Enhance the Performance of Mobile-Business Systems

[1]Wilfred W.K. Lin, [1]Allan K.Y. Wong, [2]Tharam S. Dillon and [2]Elizabeth Chang
[1]Department of Computing, Hong Kong Polytechnic University, Hong Kong SAR
[2]DEBI Institute, Curtin University, Perth, Australia
cswklin@comp.polyu.edu.hk and csalwong@comp.polyu.edu.hk

Abstract

We propose a novel wavelet-based approach (WBA) that accurately decides if the Internet traffic is currently SRD (short-range dependence) or LRD (long-range dependence) on the fly. As a result, it helps a real-time dynamic cache size tuner such as the MACSC (Model for Adaptive Cache Size Control) maintain the given cache hit ratio under all Internet traffic conditions. The WBA mechanism utilizes the following unique characteristics of wavelet coefficients (WC) for stationary signals $f(x)$: i) WC values by discrete wavelet transform (DWT), irrespective to if $f(x)$ is SRD or LRD, are always SRD; ii) if $f(x)$ is LRD then the WC variances scale divergently with the dilation factor; and iii) WC variances of SRD signals, however, converge to a constant. The DWT process is $D_{j,k} = \left\langle f(x), \omega_{j,k} \right\rangle$ or $2^{-j/2} \int_{-\infty}^{+\infty} f(x)\omega(2^{-j}x-k)dx$ formally, where $D_{j,k}$ are the WC values, $\omega_{j,k}$ the mother wavelet, j the dilation factor/scale, and k the translation factor.

Keywords: wavelet, dynamic cache size tuning, stationary signal, DWT, MACSC, WB

1. Introduction

The proposed wavelet-based approach (WBA) enhances the performance of mobile-business (MB) systems [1] by shortening the mean service roundtrip time (RTT). This makes the clients happy and return for more business. Every MB setup is a digital ecosystem of various functional species that collaborate harmoniously [2]. The inter-species interactions are in a client/server relationship via message passing. The novel WBA is suitable for MB systems with voluminous data retrievals over the Internet. Figure 1 is a model of such an MB system, exemplified by the successful telemedicine platform of the PuraPharm Group [3].

Figure 1 shows two types of MB clients: i) wireless, with SFF (*small form factor*) devices such as mobile phones (e.g. #1); and ii) wireline, such as a desktop PC (e.g. #2). Request streams from different clients converge at the SAP (*service access point*) of the intelligent proxy server (IPS). For those requests that retrieve data objects, the IPS has to search the *data/knowledge base* (D/KB) and return the found items to complete the client/server service loop (i.e. service type 1). In Purapharm's telemedicine platform [3] the requests would be part of the statistics (i.e. service type 2), which drive the drugs/herbs manufacturing process (i.e. OMA in Figure 1) adaptively on-line so that supply and demand can be quickly balanced anytime and anywhere.

The traffic pattern after convergence (i.e. "+") at the SAP is unpredictable. Its ill effects can affect the entire MB system in various forms, such as: i) buffer overflow [4]; ii) cache hit ratio reduction [5]; and iii) system failure [6]. The WBA resolves the cache hit ratio reduction problem due to Internet traffic ill effects by making the *dynamic cache size tuner* (DCST) more accurate. For this research the DCST used for experiments is the extant MACSC (Model for Adaptive Cache Size Control) [7], which strives to maintain the given cache hit ratio h under all Internet conditions.

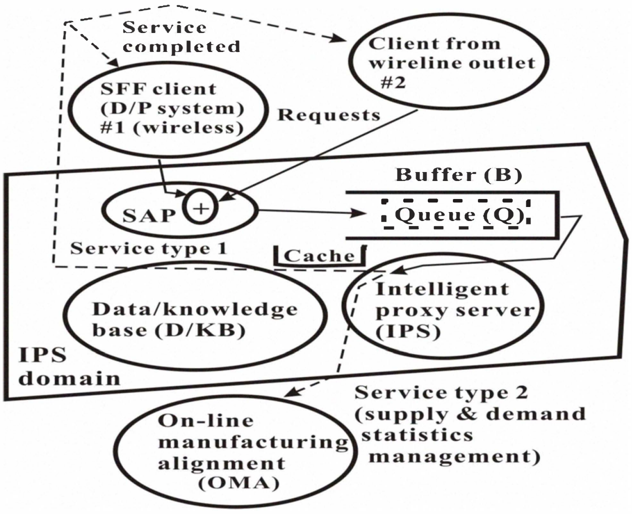
Service completed
Client from wireline outlet #2
SFF client (D/P system) #1 (wireless)
Requests
Buffer (B)
SAP
+
Queue (Q)
Service type 1
Cache
Data/knowledge base (D/KB)
Intelligent proxy server (IPS)
IPS domain
Service type 2 (supply & demand statistics management)
On-line manufacturing alignment (OMA)

Figure 1. A concise mobile business (MB) model

Effective caching helps the success of Internet-based systems [5] because it shortens the mean client/server service RTT. But, the cache hit ratio h may fluctuate over time for various reasons such as Internet traffic pattern changes (e.g. inter-arrival times (IAT) among requests). Traffic-based dynamic cache tuning maintains h by two real-time steps that form one control cycle: i) identifying the current traffic pattern (i.e. non-linear, stationary, SRD, and/or LRD [4]); and ii) tuning the cache size for the current traffic pattern detected. The proposed WBA makes these two steps accurate.

2. Related Work

The success of any MB system hinges upon making clients happy and return customers, and this is tied in with quick service response (i.e. short service RTT) [1]. In this light, efficacious caching for a high hit ratio h is the key. With this in mind a group of researchers had successfully developed the MACSC (Model for Adaptive Cache Size Control [3] for real-time applications. Simulations showed that the MACSC indeed has the ability to maintain the given h. In real-world applications, however, it was found that the h value under the MACSC control, with no WBA support, can drop seriously. As a result, the data retrieval time is lengthened tremendously. Our critical analysis of this h dipping phenomenon reveals that the cause is the sudden change in the Internet traffic pattern. Every traffic pattern has a unique complex temporal correlation; thus its time series is sensitive to a particular time scale. For example, SRD/LRD time series is sensitive to small/large time scales. This sensitivity leads to errors in the real-time statistical computation needed to confirm the current traffic pattern/characteristic/model. But, the computation accuracy may be improved if the temporal correction complexity in the target signal $f(t)$ is diluted by using wavelets [8, 9].

2.1 MACSC Definition

The MACSC is completely outlined by the equations (2.1), (2.2) and (2.3), and the relevant details are:

a) Zipf-like [10]: It is formally $y(r) \propto r^{-\beta}$ or $f(r) = \log(y(r)) = \alpha - \beta \log(r)$ alternatively for $0 < \beta \leq 1$. $y(r)$ is the access frequency for the data object of the r^{th} ranked position. In fact, the $f(r)$ linear regression shows the relative popularity among the data objects at anytime within the set. The r^{th} position of an object may change with time due to current clients' preference.

b) Zipf bell-shape $bell_i(r)$ curve [3]: Formally it is $bell(r) = map(f(r)) + e$; the map () function transforms the $f(r)$ regression into $bell_i(r)$, where the error e can be ignored because the MACSC operation is self-compensating.

c) Standard deviation (δ_i^{bell}) of $bell_i(r)$: Computed in the i^{th} dynamic cache tuning cycle.

d) Tuned/adjusted cache size (TCS_i): Computed for the i^{th} cycle by equation (2.1).

e) *Point estimate* (PE): This is formally expressed by equation (2.2), where: i) E is the specified fractional error between the true/ideal/population mean λ and the mean $\bar{r}$ of a sample of size $n \geq 10$; ii) $\delta_{\bar{r}}$ is the standard deviation of the *standard error* (SE) plot with a series of sample means or $\bar{r}$ values; iii) k is the specified number of $\delta_{\bar{r}}$ away from μ (the mean of SE) and still be tolerated as correct (same connotation as E) ; iv) δ_r is the ideal/true standard deviation of the population; and v) N is the ideal sample size that satisfies equation (2.2), which is derived from the Central Limit Theorem. In reality, one can use any sample size n to start the PE computation and repeat it by increasing n until $n \geq N$ is satisfied. For example, given $E = 0.046$ (i.e. same as $k = 2 = 95.4\% = 0.954$) and $n = 60$, we can use equation (2.3), where s_r is the standard deviation of the same sample (i.e. 60 r values) that yields $\bar{r}$. Assuming $s_r = 9$ and $\bar{r} = 15$ then equation (2.3) yields roughly $N = 680$. Since the $n \geq N$ criterion is not satisfied, more r values have to be sampled (i.e. enlarging n) and the computation repeated with equation (2.3) until $n \geq N$. If the approach $n_j = 2 * n_{j-1}$ (n_j is the sample size of the j^{th} computation trial), is adopted, the convergence to $n \geq N$ may happen in the second or third trial. The starting sample size n, however, can affect the accuracy in the maintenance of the given hit ratio h. Our critical analysis concludes that this is caused by the delay in adjusting the cache size in a timely fashion on the fly. That is, when the MACSC is computing the next fittest TCS_i value, the system is still working with the outdated cache size. To resolve this delay problem, it is necessary to help the system decide quickly what n value to use when executing equation (2.3) in the current i^{th} cache tuning cycle. In the WBA we resolve the "*what n value to use*" problem by calibration, and the problem of "*timeliness*" by identifying the current traffic pattern (i.e. SRD or LRD).

$$TCS_i = CacheSize_{i-1}\left(\frac{\delta_i^{bell}}{\delta_{i-1}^{bell}}\right)...(2.1)$$

$$E\lambda = k\delta_{\bar{r}} = k\left(\delta_r \Big/ \sqrt{N}\right)...(2.2)$$

$$N = \left(\frac{k\delta_r}{E\lambda}\right)^2 \approx \left(\frac{ks_r}{\bar{E}r}\right)^2,\ n \geq N...(2.3)$$

2.2 Discrete Wavelet Transform

A mother wavelet ω, an analysis function or a window, can be expanded/dilated/scaled to create a wavelet family in the normalized form $\omega_{a,b}(x) = \omega \Big/ \sqrt{a}\left((x-b)\Big/ a\right)$, where a and b are the scale and translation factors respectively. Changing a and/or b creates self-similar patterns. In the translational action each wavelet is localized at $x = b$ when moving along the x axis. To create the orthonormal basis we set $a = 2^j$ and $b = 2^j k(j,k \in Z)$, and change $\omega_{a,b}(x)$ to $\omega_{j,k}(x) = 2^{-j/2}\omega(2^{-j}x - k)$. For $f(x)$ the discrete wavelet transform (DWT) is $D_{j,k} = \sum_{x=0}^{2^K-1} f(x)\omega_{j,k}(x)$ for $0 \leq x < 2^K$. From the $D_{j,k}$ wavelet coefficients the original $f(x)$ can be reconstructed by inverse wavelet transform [9]. In continuous time the DWT process is $D_{j,k} = \langle f(x), \omega_{j,k}\rangle$ or $D_{j,k} = 2^{-j/2}\int_{-\infty}^{+\infty} f(x)\omega(2^{-j}x - k)dx$ conceptually. For any chosen (j,k) (i.e. scale, translation) pair, a unique $D_{j,k}$ wavelet coefficients sequence can be obtained.

The $D_{j,k}$ distribution is a representation of the original stationary signal $f(x)$ but of *diluted* temporal correlation complexity. Thus, $D_{j,k}$ is always SRD (short-range dependence) even when $f(x)$ is LRD (long-range dependence) [9]. This peculiar property is useful for the novel wavelet-based approach (WBA) proposed in this paper, for: i) reduced temporal correlation complexity means more accurate statistical computations; ii) variances δ_j of the $D_{j,k}$ sequences derived from a SRD $f(x)$ with varying j converge to a constant C (i.e. $\lim_{j\to\infty}\delta_j \to C$); and iii) if $f(x)$ is of the LRD nature, then $\lim_{j\to\infty}\delta_j \to \infty$. This means that for real-time applications we can dilute the temporal correlation complexity of a time series $f(x)$ to achieve more accurate statistical computations, and then decide quickly if the $D_{j,k}$ sequences came from a SRD or LRD $f(x)$ by checking $\lim_{j\to\infty}\delta_j \to C$ or $\lim_{j\to\infty}\delta_j \to \infty$ on the fly respectively. $\lim_{j\to\infty}\delta_j \to \infty$ holds for LRD because of $\delta_j \approx 2^{j(2H-1)}(2^{2(1-H)} - 1)\sigma$ [11]. σ is the variance of $f(x)$ in the Hurst (H) value range $(0.5 < H < 1)$ indicating SRD. $\delta_j \propto j$ holds (i.e. δ_j grows with the scale j) because the $(2^{2(1-H)} - 1)\sigma$ term is a constant. If $f(x)$ is SRD, then the $D_{j,k}$ values are Gaussian variables with a zero mean and a variance δ_j. In fact, δ_j is the sum of the following two terms: i) $\sigma\{1 + 2\rho/(1-\rho) - 3\rho/[(1-\rho)^2 2^{j-1}]\}$; and ii) $O(\rho^{2j-1})$, where ρ is a fractional constant and $O(\rho^{2j-1})$ is the remainder. For large j scales, both factors, $3\rho/[(1-\rho)^2 2^{j-1}]$ and $O(\rho^{2j-1})$ converge to zero. Thus, $\lim_{j\to\infty}\delta_j \to \sigma\{1 + 2\rho/(1-\rho)\}$ converges to a constant C.

3. The Proposed Wavelet-Based Approach

The proposed wavelet-based approach (WBA) adjusts the initial sample size n when equation (2.3) is to be executed on the fly. Theoretically, a real-life stationary time series embeds both SRD and LRD segments in an interleaved fashion. If the condition of $n \geq N$ is quickly attained, then the chance for the cache hit ratio to dip is reduced because adjusted cache size or TCS_i (i.e. equation (2.3) is quickly computed. For a stationary signal $f(x)$, the WBA detects if the $\lim_{j\to\infty}\delta_j \to C$ condition holds consistently. If true, $f(x)$ is SRD; otherwise it would be LRD. That is, if $\log_2(\delta_j)$ is plotted against different j values, it should converge to $C \approx \sigma\{1+2\rho/(1-\rho)\}$ conceptually. Therefore, $R = \lim_{j\to l}\left(\delta_j / \delta_{j+1}\right) \to 1$ is the WBA principle to detect the convergence for SRD. In light of the objective function $\{1,\Delta\}^2$ convergence is attained for $|1-R| \le \Delta$, another way of looking at this principle.

4. Simulation Results

The setup for the simulations is shown in Figure 2. The same stationary signal $f(x)$ or time series is fed into the DWT and WBA mechanisms. The DWT generates the variances δ_j of the $D_{j,k}$ wavelet coefficients for $f(x)$ by using different j, and the WBA decides if $f(x)$ is SRD and LRD. For a simulation, the input $f(x)$ may be purely SRD, LRD or an interleaved version of them. If the WBA mechanism has encountered a LRD segment in the input $f(x)$, then the initial sample size $n2$ will be suggested to the MACSC mechanism for computing equation (2.3), which precedes computing TCS_i by equation (2.1). If the current $f(x)$ segment is SRD, $n1$ is suggested to MACSC instead conceptually. Many simulations were conducted, and the results unanimously indicate that the proposed WBA mechanism helps MACSC maintain the given cache hit ratio h as the minimum in a consistent manner. This consistency is shown by the verification result in Figure 3. The difference in the two hit-ratio curves shown in Figure 3 is that one curve was obtained with WBA support, while the other is the conventional MACSC operation without such support. When WBA is absent the MACSC would use any reasonable sample size n to compute equation (2.3) and keep on using the same n if repeated computations are needed. With WBA support, however, the MACSC would use different calibrated n values under different Internet traffic conditions, such as $n1$ and $n2$ shown in Figure 2. Figure 4 shows how such calibrations can be carried out.

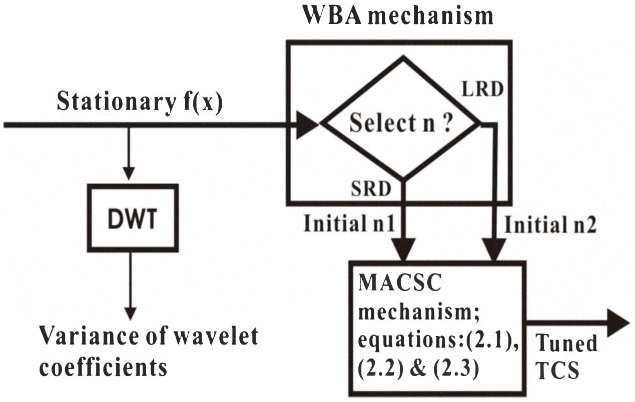
WBA mechanism
Stationary f(x)
Select n ?
LRD
SRD
DWT
Initial n1
Initial n2
MACSC mechanism; equations:(2.1), (2.2) & (2.3)
Tuned TCS
Variance of wavelet coefficients

Figure 2. Setup for the simulation experiments

4.1 Better Cache Hit Ratio Maintenance with WBA

In these simulations the two datasets, namely, $f_1(x)$ and $f_2(x)$ were arbitrarily interleaved to form different resultant signals (e.g. $f_3(x)$, $f_4(x)$,....). Then, these resultant signals were put together with some pre-collected real-life traffic traces [12], as well as a set of artificially generated traces using the MATLAB package, to form the excitation set (ES). Each trace in ES (i.e. f_D) would serve as the input RTT to drive the MACSC simulation in turn. The aim is to profile how the novel WB approach (or WBA) proposed in this paper would enhance the MACSC cache hit ratio maintenance process. We expected that the MACSC with WBA support would maintain the given hit ratio of $h = 0.643$ (i.e. 64.3% or 1 standard deviation of confidence) as the minimum value in a consistent fashion.

In the WBA-supported MACSC simulations, the setup is based on the previous experience [3], with basic support as follows: i) LRU (*least frequently used*) replacement strategy for flushing the cache; ii) TCS_i computation by equation (2.1); and iii) f_D, which has an average request arrival rate (i.e. $1/\left(\overline{IAT}\right)$), becomes the request IAT (inter-arrival time) function to drive the data retrieval service request rate (i.e. conceptually the stationary $f(x)$ in Figure 2). The traffic pattern together with the average request arrival rate at any time would affect the cache hit ratio of the IPS (Figure 3). The MACSC mechanism would compute TCS_i (i.e. equation (2.1)) as quickly as possible in order to maintain the given cache hit ratio h, which represents the target performance of the IPS caching system.

The WBA mechanism helps ensure two objectives: i) the maintenance of the given h by the MACSC is persistent; and ii) the resultant cache hit ratio, in the sense of continuous time, is above this given h as much as possible.

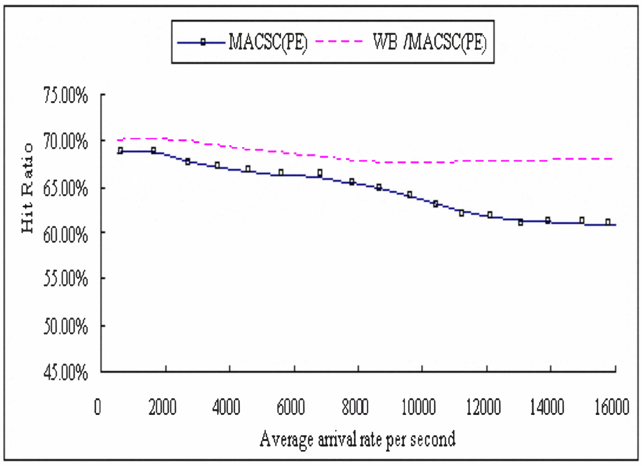

Figure 3. MACSC(PE) with MB support always yields higher cache hit ratio than the given h

Figure 3 is the plot of the hit ratio versus every mean request arrival rate or $1/\left(\overline{IAT}\right)$, where $\overline{IAT}$ is the mean IAT value embedded in f_D. Obviously, as $\overline{IAT}$ gets shorter (i.e. $1/\left(\overline{IAT}\right)$ getting higher), the *point-estimate-based* MACSC or MACSC(PE), without WBA support, would take longer to compute equation (2.3) and satisfy the $n \geq N$ criterion. The delay causes the hit ratio to drop below the expected/target value of $h = 0.643$. The WBA presence rectified the hit-ratio dropping situation by deciding if the MACSC mechanism should use $n1 = 100$ for SRD at the time or switch to $n2 = 128$ immediately, or vice versa. Since this WBA decision saves the trial time for computing the $n \geq N$ criterion, the MACSC can tune the cache size more proactively to maintain at least the given h cache hit ratio. The simulation result in Figure 3 shows that the hit ratio produced by the MACSC dynamic cache size tuning mechanism, supported by the novel WBA mechanism, is consistently above the given $h = 0.643$. In fact, other similar simulations with different sets of f_D distributions produced the same phenomenon. That is, the MACSC, if supported by the novel WBA mechanism, would maintain the given h as the minimum.

4.2 Sample Size Calibration

The sample sizes for the MACSC to use under different traffic conditions have to be obtained by calibration. For example, in the simulation that produced Figure 3 the sample size for the SRD conditions is $n1 = 100$, and for the LRD conditions it is $n2 = 128$. The calibration setup, which has no WBA support, is similar to Figure 2, and the calibration steps are as follows:

a) Firstly, a time series (or $f_1(x)$) of the SRD nature is used to excite the MACSC, which works with an arbitrarily chosen n to compute equation (2.3) and the given hit ratio to be maintained is $h = 0.643$ (i.e. 1 standard deviation). Then, different n values should be tried out to produce different distributions of hit ratios. After that, the same process is repeated by using different time series $f_k(x); k = 1,2,...j$. Finally all the hit ratios with respect to the specific n value (sample size) and different $f_k(x)$ are averaged to get the calibration point. In this calibration exercise the best n that has yielded the highest average hit ratio is the calibrated $n1$ for SRD; $n1 = 100$ for the experiment that produced Figure 3.

b) Secondly, different LRD time series are used to repeat the calibration procedure as for yielding $n1 = 100$. In light of Figure 3, it is $n2 = 128$.

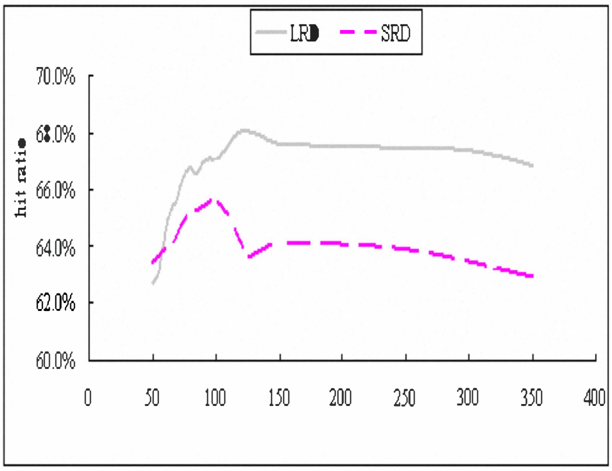

Figure 4. Calibration "hit ratios vs. n values" for use by equation (2.3); $n1 = 100$ for SRD & $n2 = 128$ for LRD

Figure 4 summarizes SRD and LRD calibrations as the prelude to the simulation that produced the data in Figure 3. In this case, the samples sizes for SRD and LRD traffics are respectively $n1 = 100$ and $n_2 = 128$.

5. Conclusion

In this paper we propose the wavelet-based approach (WBA), which makes use of the discrete wavelet transform (DWT) that: a) dilutes the temporal correlation complexity of stationary signal $f(x)$; and b) yields different behavior for the variance δ_j of wavelet coefficients with respect to the scale parameter j. That is, i) for a SRD $f(x)$, $\lim_{j\to\infty}\delta_j \to C$ where C is a constant, and ii) for a LRD $f(x)$, $\lim_{j\to\infty}\delta_j \to \infty$. Once the SRD or LRD nature is confirmed, the WBA would suggest $n1$ (for SRD) or $n2$ (for LRD) as the initial sample size for the MACSC to execute equation (2.3) so that the criterion of $n \geq N$ can be quickly satisfied on the fly. This quick satisfaction holds the key to the MACSC(PE)'s ability in maintaining the given hit ratio h as the minimum value. As a result of this ability, the mean client/server service roundtrip time (RTT) is persistently shortened, leading to fast system response that makes mobile-business (MB) customers happy and return for more business. In this light, the contribution by the novel WBA proposal is significant. The next two logical steps in this research are as follows:

a) *Quicker confirmation*: The aim is to explore how to confirm the $R = \lim_{j\to l}\left(\delta_j / \delta_{j+1} \right) \to 1$ criterion more quickly on the fly. A quicker confirmation as such would make the novel WBA mechanism even more suitable for real-time applications in general.

b) *Automatic sample size calibration*: Since the accuracy of $n1$ and $n2$ would affect the speed and accuracy of the N (equation (2.3) and TCS_i (equation 2.1) computations. If the WBA can calibrate these values from the traffic conditions continuously and automatically, $n1$ and $n2$ can be adaptively changed for more accurate results.

6. Acknowledgement

The authors thank the PuraPharm Group for the research fund H-ZW93.

7. References

[1] S.F. Thomas and M.L. Gillenson, Mobile Commerce: What It Is and What It Could Be, Communications ACM, 46(12), December 2003, 33-34

[2] Digital Business Ecosystems, eds. F. Nachiro et al, European Commission, Information Society and Media, 2007

[3] Jackei H.K. Wong, Allan K.Y. Wong, Tharam S. Dillon and Wilfred W.K. Lin, Text Mining as a Technique to Support Effective TCM (Traditional Chinese Medicine) Telemedicine Evolution, in Domain Driven Data Mining: Domain Problems and Applications, Eds. Philip S. Yu et al, Springer 2008, 143-157

[4] Wilfred W.K. Lin, Allan K.Y. Wong and Tharam S. Dillon, Application of Soft Computing Techniques to Adaptive User Buffer Overflow Control on the Internet, IEEE Transactions on Systems, Man and Cybernetics, Part C, 36(3), May 2006, 397 -410

[5] J. Wang, A Survey of Caching Schemes for the Internet, ACM SIGCOMM, 29(5), 1999, 36-46

[6] V. Paxson and S. Floyd, Wide area traffic: The Failure of Poisson Modeling, IEEE/ACM Transactions on Networking, 3(3), 1995

[7] Allan K.Y. Wong, May T.W. Ip and Richard S.L. Wu, A Novel Dynamic Cache Size Adjustment Approach for Better Data Retrieval Performance over the Internet, Computer Communications, vol. 26, 2003, 1709-1720

[8] N.E. Miner, An Introduction to Wavelet Theory and Analysis, Sandia National Laboratories, SAND98-2265, 1998

[9] S. Ma and C. Ji, Modeling Heterogeneous Network Traffic in Wavelet Domain, IEEE/ACM Transactions on Networking, 9(5), 2001

[10] L. Breslau et al., Web Caching and Zipf-like Distribution: Evidence and Implications, Proc. of the IEEE INFOCOM, 1999

[11] L.M. Kaplan and C.J. Kuo, Fractional Estimation from Noisy Data Via Discrete Fractional Gaussian Noise (DFGN) and the Haar Basis, IEEE Transactions on Signal Processing, vol. 42, 1993, 3554 – 3562

[12] The Internet Traffic Archive, ACM SIGCOMM Special Interest Group on Data Communications, http://ita.ee.lbl.gov/index.html