Adaptive Transmission Scheme for Web Prefetching in Wireless Environment

Ryoichi SHINKUMA\textsuperscript{[a]}, Student Member, Minoru OKADA\textsuperscript{[†]}, and Shozo KOMAKI\textsuperscript{[††]}, Regular Members

SUMMARY This paper proposes an adaptive transmission scheme for web prefetching in wireless communication systems. The proposed adaptive transmission scheme controls the modulation format and the error control scheme according to the access probability of the web document being transmitted. In the proposed system, the actually requested documents and the documents which have high access probability are transmitted with a reliable transmission format, while the pages whose access probabilities are lower than a certain threshold are transmitted with a bandwidth efficient transmission format. The computer simulation results show that the proposed scheme drastically improves the latency performance.

\textit{key words:} wireless communications, adaptive transmission, world wide web, prefetch

1. Introduction

The world wide web has become a killer application not only of the wired internet but also of the cellular mobile systems. However, since the radio resources are strictly limited, the growth of mobile web users leads the degradation in the web document access latency performance. Web prefetching scheme \cite{1}–\cite{4} has been developed to overcome the problem. Since the user usually spends some time for viewing a page, the channel is idle during this period. In the prefetching scheme, the pages that may be requested soon are downloaded in such an idle time. If the user requests a page that is already fetched during the idle time and stored in the local storage, there will be no transmission delay for the user. This implies that the prefetching scheme can improve the web access latency performance. In the wireless web access, we can improve the latency performance by using the prefetching scheme \cite{4}. However, the frequency utilization efficiency decreases as increase in the number of prefetched pages, since the prefetch scheme wastes the bandwidth when the most prefetched pages are not actually requested by the user. In addition, although the transmission power is wasted as well as the bandwidth, it is not so serious problem since all of prefetched data are transmitted from the base station in the downlink.

In order to improve the frequency utilization efficiency, one would propose a system that makes use of bandwidth efficient modulation formats such as 16QAM (Quadrature Amplitude Modulation) and 64QAM. These modulation formats can realize high speed transmission in the narrow bandwidth channels. However, since the bandwidth efficient QAM schemes are more sensitive to disturbances due to additive noise and fading, we cannot use them for the transmission requiring high reliability.

Recently, the adaptive transmission methods for wireless multimedia networks have been researched \cite{5}–\cite{12}. Some adaptive transmission methods adaptively change the service quality on application layer according to the physical parameters such as the channel CNR (Carrier to Noise power Ratio), available bandwidth or capability of the user terminal \cite{5}–\cite{8}. However, these types of adaptive transmission systems require the feedback channel to send channel state information from the terminal to the base station in order to track the channel state. Furthermore, we can not apply it to the transmission for multimedia data including text data although it is possible to change the service quality by controlling the compression rate for the voice data or image data. The other adaptive transmission methods adaptively control the transmission formats, such as the modulation format and error control scheme, according to the parameters on the application layer, instead of physical parameters. Reference \cite{9} shows the adaptive radio design, which decides the frame length, error control and bit rate based on QoS (Quality of Service) level indicated by the QoS manager on the application layer. Although \cite{9} assumes that the QoS parameters consist of the throughput, delay and error rate, it doesn’t discuss difference of the QoS level required by the data whose formats are different each other. On the other hand, a part of the authors has proposed the efficient image transmission scheme for the mobile communication environments \cite{11}. It first divides image data into several groups according to their error sensitivity and transmits each group with different modulation format. Furthermore, \cite{12} has proposed the mixed media transmission scheme, which chooses modulation level depending on required QoS level such as bit er-
ror rate and transmission delay. Although the schemes proposed in [11] and [12] can improve the total QoS and the utilization of the radio resources by adaptively controlling the transmission format according to the QoS level required by the data, it can not directly be applied to the wireless web browsing, since it employs the adaptive modulation only for image transmission, which is insensitive to bit errors.

In this paper, we propose a novel adaptive transmission scheme, which is capable of improving the bandwidth utilization efficiency of the web prefetching scheme in a wireless communication system. The proposed scheme chooses the transmission format for the prefetched pages according to their access probabilities. When the proposed system sends the prefetched page to the user terminal, the reliable modulation format such as QPSK (Quadrature Phase Shift Keying) will be employed if its access probability is higher than a threshold. On the other hand, prefetched pages with lower access probability are transmitted with the bandwidth efficient modulation schemes. As the data link error control method suitable to the transmission for web documents, we employ the ARQ (Automatic Repeat reQuest) error control in the proposed scheme. In order to improve the frequency utilization efficiency, the proposed scheme adaptively performs ARQ error control according to the access probability when the prefetched data are transmitted.

In wired network techniques, a number of QoS control schemes on the data link layer or IP (Internet Protocol) layer has been proposed [13],[14]. These QoS techniques mainly use scheduling methods in order to match the delay and jitter requirements. For instance, the high priority is given to the delay sensitive data such as streaming video and the low priority is given to the delay insensitive data such as prefetched data. However, since the bandwidth in the wireless channel is strictly limited and the error rate performance of the wireless link is much worse than that of the wired one, we have to determine a new mapping scheme that maps the QoS priorities onto wireless link in order to match the QoS requirement and the wireless channel characteristics. This paper focuses on the data link QoS control and the adaptive transmission scheme for the wireless web access. The rest of the paper is organized as follows. Section 2 proposes the adaptive transmission scheme, followed by the computer simulation results shown in Sect. 3. Finally, concluding remark is given in Sect. 4.

2. System Description

Figure 1 illustrates the block diagram of the proposed adaptive transmission scheme, which is composed of the gateway server, base station and mobile terminal. The prefetch agent on the gateway server fetches the web data and gives the QoS (Quality of Service) label to each of packet data. The QoS manager decides the transmission format according to the QoS label. The adaptive transmitter of the base station transmits data with the transmission format decided by the QoS manager on the gateway. The terminal prefetch agent predicts the access probability of the page and requests the data to the gateway prefetch agent.

2.1 Adaptive Transmission Scheme

Let us go into further detail of the proposed scheme. Figure 2 and Table 1 show the adaptive transmission procedure performed on the gateway server and base station. The prefetch agent sends data requested by the user to the QoS manager with QoS label A immediately. QoS manager controls the data link transmission format, namely, the modulation format as well as the maximum number of retransmission on ARQ. When QoS manager catches the label A packet, it chooses QPSK (Quadrature Phase Shift Keying) and unlimited ARQ to transmit that packet. The unlimited ARQ performs retransmission till the transmitted packet is received correctly. On the other hand, web documents that should be requested are prefetched and

Table 1 Transmission scheme selection rule.

<table>
<thead>
<tr>
<th>Access prob.</th>
<th>QoS label</th>
<th>Mod. format</th>
<th>ARQ scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requested</td>
<td>A</td>
<td>QPSK</td>
<td>unlimited</td>
</tr>
<tr>
<td>$P &gt; P_{th}$</td>
<td>B</td>
<td>QPSK</td>
<td>limited</td>
</tr>
<tr>
<td>$P \leq P_{th}$</td>
<td>C</td>
<td>2^p-QAM</td>
<td>limited</td>
</tr>
</tbody>
</table>
stored on the storage of the gateway server. When all of requested data has been transmitted, the gateway prefetch agent starts transmitting the prefetched data. In this process, prefetch agent gives the QoS label to the prefetched data according to the access probability. If the access probability of the transmitted data is higher than the threshold $P_{th}$, it will attach the label $B$ to the data. While, the label $C$ is given to the data whose access probability is lower than $P_{th}$. QoS manager employs QPSK modulation for the label $B$ packet, while it uses 16QAM (Quadrature Amplitude Modulation), 64QAM or 256QAM for label $C$ packet. As the error control scheme for the label $B$ or $C$ packets, the limited ARQ scheme is proposed. The limited ARQ scheme restricts the maximum number of retransmission per one packet. In case that many error packets are received while a prefetched page is transmitted and the number of retransmission exceeds the limit, the adaptive ARQ will stop transmitting the page and starts transmitting the next prefetched document. Successfully transmitted prefetched documents are stored on the local storage. If the user requests the page that has been fetched and transmitted in advance, the browser can show the contents of the page to the user soon. Otherwise, the terminal prefetch agent asks the gateway prefetch agent to stop prefetching and to fetch the requested data from the web server.

2.2 Block Diagrams

In order to realize the adaptive transmission procedure described in 2.1, the base station and terminal are composed as follows. Figure 3(a) shows the gateway prefetch agent, the QoS manager and the adaptive transmitter. The prefetch agent consists of the prefetch manager and cache. QoS manager has the packet processing unit and QoS controller. The adaptive transmitter has the $2^k$-QAM modulator, ARQ manager and adaptive controller. Figure 3(b) shows the terminal prefetch agent and adaptive receiver. The prefetch manager and prediction module run on the terminal prefetch agent. The adaptive receiver consists of the $2^k$-QAM demodulator, modulation level controller and ARQ manager. In the following, we give more details of the adaptive transmission procedure.

2.2.1 Transmission Procedure

In the terminal, the prefetch agent first estimates the probability that a prefetched page is actually requested by the user. Then the prefetch manager requests the data of the prefetched page and transmits the value of its user’s access probability estimated by the prediction module to the gateway prefetch agent. On the other hand, on the gateway server, the prefetch manager of the gateway prefetch agent fetches the data of the page from the WWW server and stores it on the cache. And the prefetch manager gives the QoS label $B$ or $C$ each packet of the data according to the value of user’s access probability transmitted by the terminal. The QoS controller decides the modulation level of the $2^k$-QAM modulator according to the QoS label. The packet processing unit adds the data link header to the packets and encodes them with the error detecting code for ARQ error control. The adaptive controller makes the $2^k$-QAM modulator transmit the data with the modulation format indicated by the QoS controller. In this process, if the transmitted packet label is $B$, QPSK is employed. On the other hand, if label is $C$, it is transmitted by 16, 64 or 256 QAM.

2.2.2 Reception Procedure

The terminal adaptive receiver demodulates the received signals from the base station according to the modulation type shown by the base station. The ARQ manager gets the demodulated packets and then controls errors by making use of the ARQ procedure. The ARQ manager checks errors in the received packet, and ACK (ACKnowledgement) signal is returned to the base station and the packet data are stored on the cache if the packet is transmitted correctly. On the other hand, if the packet has error bits, the ARQ manager returns NAK (Negative ACK) to request the base station for retransmitting the packet. In case that the ARQ manager of the base station catches a lot of NAK signals while the label $B$ or $C$ packet is transmitted, the ARQ manager sends the error message to the QoS controller. Then, the QoS controller asks the prefetch...
manager to stop transmitting the current prefetched page and start transmitting the next page. Thus, the user can browse the page immediately if the terminal has received all data of it. Otherwise, the prefetch manager must ask the gateway server to transmit the necessary data that have not been transmitted yet. Since the necessary data is cached in the gateway prefetch agent, the prefetch manager resumes transmission as soon as the user requests the page. The data actually requested by the user is given label A and transmitted reliably with QPSK and the unlimited ARQ error control. Finally, the WWW browser shows the contents of the page when the terminal has received its all data perfectly.

3. Performance Analysis

In this section we analyzed the latency performance of the proposed scheme in a mobile communication environment by computer simulation. For the sake of simplicity, we assume the point-to-point connection between one user terminal and one base station. The system operates in TDD (Time Division Duplex) mode. Although web documents are sent from web servers on the internet to the mobile terminal through both the wired and wireless channels, the delay in the wireless channel was only taken into account for analyzing the latency performance since the bandwidth of the wireless channel is usually much narrower than that of the wired channel. Let us define the idle time between requests, or the interval between the end of transmission of a requested web document and the next request, as \( \Delta t \). In the simulation model, WWW pages consist of a text file and a still image files. The size of text and image files is modeled as a lognormally distributed random variable as in [16]. The proposed scheme has to estimate the request probability. Some algorithms for the access probability prediction has been proposed in [2]–[4]. For example, the algorithm proposed in [3] and [4] estimates the access probability from the user’s browsing patterns. Although our proposed scheme in fact has to employ such a prediction algorithm, in order to simplify the simulation, we assumed that all the pages linked to the already fetched page have an equal probability. For example, when the fetched page has \( m \) linked pages, the probability of each linked page is \( 1/m \). We assume the flat Rayleigh fading channel, and the receiver has the perfect knowledge of the channel state. Table 2 shows the system configuration of the simulation.

The proposed scheme employs QPSK for transmitting the prefetched pages whose access probabilities are more than \( P_{th} \), while it employs 16 QAM for transmitting the prefetched pages whose access probabilities are less than \( P_{th} \), where \( P_{th} \) is the threshold for determining the modulation levels. If retransmission for one packet is performed \( N \) times while a certain page is prefetched, transmitting the page is stopped and then start prefetching the next page. Since the data that the user actually requests are very important, they are fetched with the QPSK modulation format and the unlimited ARQ scheme. In fact, since the total performance of the communication system may be degraded if the number of retransmission is not restricted, we must set the timer for the ARQ. However, in order to evaluate the performance of the proposed system only based on the latency, we do not deal with the timeout here.

### 3.1 Distribution of Latency Time

Now we show the latency performance of the proposed scheme. Figure 4 shows the probability that the latency is less than abscissa, namely, the cumulative distribution function (cdf) of latency in web browsing, in a Rayleigh fading channel. We assume the threshold of 1, 1/25 and 1/49. For comparison purposes, the latency performance of the prefetching scheme without adaptive transmission as well as of the conventional scheme without prefetching. Figures 4(a) and (b) represent the performance at CNR = 30 dB and 15 dB, respectively. In Fig. 4(a), although the proposed scheme can improve the latency performance, the conventional 16 QAM transmission with prefetching gives the best performance. For example, the probabilities that the latency is less than two seconds for the proposed scheme at \( P_{th} = 1, 1/25, \) and \( 1/49 \) are 85.1%, 82.8% and 80.0% respectively, while those for the conventional QPSK with and without prefetching, and 16 QAM with and without prefetching, are 78.7%, 57.5%, 89.8%, and 71.4% respectively. This is because the bit error rate of 16 QAM at CNR = 30 dB is low enough to transmit the actually requested data. On the other hand, in Fig. 4(b), the proposed scheme gives better performance at CNR = 15 dB. In this case, the probabilities that the latency is less than two seconds for proposed scheme at \( P_{th} = 1, 1/25, \) and \( 1/49 \) are 69.4%, 70.2% and 71.2% respectively while those for the QPSK with
Fig. 4 The probability that the latency is less than abscissa in a Rayleigh fading channel. In the proposed scheme, the performance at $P_{th} = 1, 1/25,$ and $1/49$ are evaluated. The latency of QPSK with and without prefetching as well as 16QAM with and without prefetching are plotted on the figure.

and without prefetching, and 16 QAM with and without prefetching are 71.4%, 49.6%, 55.6% and 34.7%, respectively. That is, when the CNR becomes low, the proposed scheme can improve the latency performance, since the error rate performance of the 16 QAM at CNR = 15 dB is not good enough to transmit the data actually requested by users. In this figure, the latency performance depends on threshold $P_{th}$ as well. The optimum threshold in terms of the latency performance is $1/49$ at CNR = 15 dB.

3.2 Latency Performance versus Mean CNR

Figure 5 shows the latency performance versus the mean channel CNR. The ordinate represents the probability that the latency is less than one second. In this figure, the performances of proposed scheme in case of $P_{th} = 1, 1/25$ and $1/49$ are shown as well as that of the conventional scheme with QPSK and 16 QAM. When the mean channel CNR is higher than 20 dB, Conventional 16 QAM gives best performance. However, its degradation is very large at the CNR = 15 dB. This is because the error performance for 16 QAM becomes worse in the low CNR region. Since Conventional 16 QAM transmits all data with 16 QAM, many packet errors lead to the degradation of the latency performance. On the other hand, in case of $P_{th}$ is 1 or $1/25$, the proposed scheme gives better performance than Conventional QPSK when CNR is higher than 20 dB. At CNR = 15 dB, the performance of the proposed scheme is worse than that of Conventional QPSK. However its degradation is relatively small. In case that the threshold $P_{th}$ is $1/49$, the performance is slightly better than that of Conventional QPSK independent of the channel CNR.

3.3 Latency Performance versus $P_{th}$

In the previous section, we verified that the proposed scheme gives better performance than the conventional scheme, which transmits the prefetched data and requested data with the same transmission format. However, if $P_{th}$ is inappropriate, we can not improve the latency performance. Therefore, we have to decide the optimum value of the threshold $P_{th}$.

Figure 6 shows the latency performance of the proposed scheme versus the threshold $P_{th}$. The proposed scheme employs QPSK for the prefetched page whose access probability is higher than $P_{th}$. On the other
hand, the prefetched page is transmitted with 16QAM if its access probability is lower than $P_{th}$. The performances at mean channel CNR = 15, 20, 25 and 30 dB are shown. From this figure, we can improve the latency performance with increasing the value of $P_{th}$ when the CNR is higher than 20 dB. Since 16 QAM can make efficient use of the frequency resource in the region mean CNR is higher than 20 dB, the latency performance is improved by increasing the number of prefetched pages that are transmitted with 16 QAM. On the other hand, at the mean CNR = 15 dB, the latency performance becomes worse with increasing the value of $P_{th}$. This is because the error rate of 16 QAM is high at CNR = 15 dB. Thus, the optimum value of $P_{th}$ varies according to the channel CNR. However, by setting $P_{th}$ to 0.04–0.08, the proposed scheme gives relatively the good performance regardless of the channel CNR.

4. Conclusion

In this paper, we have proposed a new adaptive transmission scheme for web prefetching in wireless communication systems. Our proposed system chooses the modulation format adaptively depending on the access probabilities. It also employs the efficient error control scheme for prefetching. We have analyzed the latency performance of the proposed scheme. From computer simulation results, we have verified that our scheme gives much better performance than conventional one when the threshold $P_{th}$ and the modulation format are properly chosen. As the next investigation, we have to research the optimized value of $P_{th}$. In the simulation analysis, we have shown the appropriate value of threshold $P_{th}$ on the assumption that the access probability of the prefetched pages is equal to each other. References [3] and [4] propose the prediction algorithm to estimate the access probability and describe the method for optimization of the threshold in the prefetching scheme. We have to analyze the performance in case of applying these algorithms to the proposed scheme. Furthermore, we have only analyzed the performance of the proposed system in the case of the point-to-point connection in this paper. We have to extend the proposed prefetch scheme into the multiple access environments. In multiple access environments, applying the prefetching scheme for one user affects the latency performances to other users because many users share the bandwidth. When someone uses a lot of radio resources for prefetching, the resource of the other user decrease and this implies the degradation in latency performance of other users. In order to solve this problem, the proposed scheme requires the efficient scheduling method, which gives low priority to prefetched data. In addition, if many users request the web data to the gateway server of the proposed system at the same time, the gateway server becomes heavy loading. On the other hand, applying the proxy caching scheme [15] and proxy prefetching scheme [4] to the gateway server allows us to get the remarkable improvement in the latency performance. Thus, in order to evaluate the performance of the proposed scheme in the multi user environments, we should drastically change the analytical model.

References


Ryoichi Shinkuma was born in Osaka, Japan on October 3, 1976. He received the B.E. and M.E. degrees in Communication Engineering from Osaka University, Osaka, Japan, in 2000 and 2001, respectively. He is currently pursuing the Ph.D. degree at Osaka University. He is engaging in the research on multimedia wireless communication systems.

Minoru Okada was born in Tokushima, Japan, on March 4, 1968. He received the B.E. degree in communication engineering from University of Electro-Communications, Tokyo, Japan, in 1990, and the M.E. and Ph.D. degrees in communication engineering from Osaka University, Osaka, Japan, in 1992 and 1998, respectively. He is currently an associate professor in the Graduate School of Information Science, at Nara Institute of Science and Technology. His current interest is in personal, indoor and mobile radio communications systems. He is a member of IEEE.

Shozo Komaki was born in Osaka, Japan, in 1947. He received B.E., M.E. and Ph.D. degrees in Electrical Communication Engineering from Osaka University, in 1970, 1972 and 1983 respectively. In 1972, he joined the NTT Radio Communication Labs., where he was engaged in repeater development for a 20-GHz digital radio system, 16-QAM and 256-QAM systems. From 1990, he moved to Osaka University, Faculty of Engineering, and engaging in the research on radio and optical communication systems. He is currently a Professor of Osaka University. Dr. Komaki is a senior member of IEEE, and a member of the Institute of Television Engineers of Japan (ITE). He was awarded the Paper Award and the Achievement Award of IEICE, Japan in 1977 and 1994 respectively.