Content Consistency Model for Pervasive Internet Access*

Chi-Hung Chi¹, Lin Liu¹, Choon-Keng Chua²
¹ School of Software, Tsinghua University, Beijing, China
² School of Computing, National University of Singapore, Singapore
Contact email: chich hung@mail.tsinghua.edu.cn

ABSTRACT. In this paper, we propose a new content consistency model for pervasive Internet access. We argue that content retrieved over the Internet consists of not only the data object but also its attributes needed to perform appropriate network or presentation related functions such as caching, content reuse, and content adaptation. With this model, four types of content consistency are defined. To get a deeper insight on the current situation of content consistency over Internet, real content on replica / CDN (Content Delivery Network) was monitored and analyzed. Surprisingly, we found that there are lots of discrepancies in data object and attributes found by comparing the original copy and the retrieved copy of the content. This result is important because they have direct implications to the trustworthiness of information over the Internet.

I. INTRODUCTION
Web caching is a mature technology to improve the performance of web content delivery. To reuse a cached content, the content must be *bit-by-bit equivalent* to the origin (known as data consistency). However, since the internet is getting heterogeneous in terms of user devices and preferences, we argue that traditional data consistency cannot efficiently support pervasive Internet access. There are two problems that have not yet been addressed: i) correctness of function execution in network, and ii) reuse of pervasive content. On Internet, there lies a fundamental difference between “data” and “content”. Data usually refers to entity that contains a single value, for example, in computer architecture each memory location contains a word value. On the other hand, content (such as a web page) contains more than just data; it also encapsulates *attributes* to administrate various functions of content delivery. Unfortunately, present content delivery only considers the consistency of data but not attributes. Web caching, for instance, relies on caching information such as expiry time, modification time and other caching directives, which are included in attributes

* This research is supported by the funding 2004CB719400 of China.
of web contents (HTTP headers) to function correctly. However, since content may traverse through intermediaries such as caching proxies, replicas and mirrors, the HTTP headers users receive may not be the original. Therefore, instead of using HTTP headers as-is, we question about the consistency of attributes. This is a valid concern because the attributes directly determine whether the functions will work properly and they may also affect the performance and efficiency of content delivery. Besides web caching, attributes are also used for controlling the presentation of content and to support extended features such as privacy and preferences.

Under pervasive Internet access, contents are delivered to users in their best-fit presentations (also called variants or versions) for display on heterogeneous devices [5] [6] [9]. As a result, users may get presentations that are not bit-by-bit equivalent to each other, yet all these presentations can be viewed as “consistent” in certain situations. Data consistency, which refers to bit-to-bit equivalence, is too strict and cannot yield effective reuse if applied to this pervasive environment. In contrast to data consistency, our proposed content consistency does not require objects to be bit-by-bit equivalent. This relaxed notion of consistency increases reuse opportunity, and leads to better performance in pervasive content delivery.

In this paper, we propose a new concept termed content consistency and show how it helps to maintain the correctness of functions and improve the performance of pervasive content delivery. With this model, four types of content consistency are defined. To get a deeper insight on the current situation of content consistency over Internet, real content on replica / CDN (Content Delivery Network) was monitored and analyzed. Surprisingly, we found that there are lots of discrepancies in data object and attributes found by comparing the original copy and the retrieved copy of the content. This result is very important because these findings have direct implications to the trustworthiness of information systems over the Internet. Results from this study also provide hints on how the quality of the content provided over the Internet can be improved.

2. CONTENT CONSISTENCY MODEL
In this section, we are going to propose our content consistency model for pervasive Internet access. Its system architecture, content model, and a new classification of content consistency will be given.

2.1. System Architecture
Our vision of the future content consistency is depicted in Figure 1. We begin by describing the pervasive content delivery process in three stages: server, intermediaries, and client. In stage one, server composes content by associating an object with a set of attributes. Content is the unit of information in the content delivery system where object refers to the main data such as image and HTML