Using Mobile Phones to Improve Offline Access to Online Information: Distributed Content Delivery

Jinglan Zhang, Shlomo Geva
Faculty of Science and Technology, Queensland University of Technology
Brisbane, Queensland, Australia

Andrew Trotman
Department of Computer Science, University of Otago
Dunedin, New Zealand

Abstract - **Mobile phones are now powerful and pervasive making them ideal information browsers. The Internet has revolutionized our lives and is a major knowledge sharing media. However, many mobile phone users cannot access the Internet (for financial or technical reasons) and so the mobile Internet has not been fully realized. We propose a novel content delivery network based on both a factual and speculative analysis of today's technology and analyze its feasibility. If adopted people living in remote regions without Internet will be able to access essential (static) information with periodic updates.**

Keywords
Mobile computing, content delivery, information access, knowledge dissemination.

I. INTRODUCTION

Since the early 1990s, the World Wide Web (the Web) has become indispensable for many people. The Internet has changed the way we communicate and the way we do business. However, people who live in regions where there is no Internet infrastructure cannot access the Internet in the usual way.

Mobile phones have also revolutionized our lives. Mobile phones are a far more important communication technology for people in the poorest countries than the traditional land-line telephone as most people in the world now have mobile phones[1]. Modern mobile phones are also very powerful with computing capacity similar to personal computers of only a few years ago. However, many mobile phone users cannot access the Internet due to financial or technical problems. The mobile internet has not yet been fully realized for knowledge dissemination.

Herein we propose to use the mobile phone as an end-user content delivery medium and push the contents of the Internet (more specifically, part of the Internet) to the device. Content will be delivered to the device by mobile phone service providers who will add this service to their existing communication services. Through our proposed service users who live in the most remote regions (without Internet or without convenient access to Internet) will be able to access at least the static and essential part of the web.

The research question herein is: is it feasible to use mobile phones to deliver large amount of web content so that the user can search and browse offline?

II. RELATED WORK

A content delivery network (CDN) is a network designed for optimal delivery of content [1]. CDNs generally consist of a cluster of computers containing copies of data, each placed at various points in a global network. For example, Akamai[2] has 84,000 servers deployed in 72 countries and serve content for, amongst others, Apple's iTunes. CDNs improve information access by replicating data at multiple sites and placing content as close as possible to users. Akamai mirror many static web sites on their network. When a user requests the content, it is served from the nearest mirror, rather than a far-away central server. This reduces Internet traffic and thus cost. With more replicas more users can concurrently be supported, and the lower the access time.

Duplicating Web content around the globe in a network of powerful computers has many benefits; however it is not financially viable for every content provider to build such a network. Consequently there are many commercial (and free) content delivery networks. In usage, Google has the largest, followed by Akamai [2]. Google has data centers distributed around the world for quick processing of queries and delivery of content. They automatically redirect queries from google.com to the nearest site (e.g. google.com.au) depending upon the user's geographic location.

Extending this data duplication scheme to the logical extreme, herein it is proposed that each mobile phone acts as a content delivery server in a global content delivery network, and that each phone holds the entire Web; or more practically, those static pages under the highest demand. This way each user can use their mobile phone to

[1] www.itu.int/ITU-D/ict/statistics/at_glance/KeyTelecom.html

[2] http://www.akamai.com/

access the entire static Web without the need to be connected online.

Traditional CDNs do not take advantage of the capacity of the end users' device; they are end-user-exclusive. Quite differently, the approach herein is end-user-inclusive and takes full advantage of the storage capacity and processing power of the device.

A Distributed Content Delivery Network (DCDN) incorporating end users has previously been proposed by Mulerikkal & Khalil [3], and is illustrated in Fig.1. Their approach requires end users to act as surrogate servers and to forward content to peers. This requires complex load-balancing algorithms and has associated privacy issues. Every surrogate also faces transmission delay due to the very low upload transmission rates.

III. NEW CDN ARCHITECTURE

Herein a new content delivery network is proposed. This network includes both online content delivery services (such as Akamai), and mobile phone sales stores. It is similar to that of Mulerikkal & Khalil [3] in that it uses multiple level content delivery services; however unlike that of Mulerikkal & Khalil it does not require end users to forward contents to others. Instead it uses pre-existing mobile phone retail outlets as a delivery mechanism for pre-packaged content on removable media. In doing so it combines the online advantages of the dynamic Web delivered through traditional CDNs with offline static content on the users' device to offer a flexible information delivery service that can be used both when the user is online and offline. The proposed architecture is illustrated in Fig.2.

In the Figure:

- Master CDN Servers are the first point of contact for the content provider. They are responsible for knowing the location of the content and for ensuring the CDN replicas are up-to-date at all times.
- Local CDN Servers are strategically located close to end uses (and sales agents). They store mirrors of the (geographically) requested web content and maintain information about that content.

Both Master CDN Servers and Local CDN Servers are necessary in existing CDN networks.

- Sales Agents are mobile phone retail outlets. They retail mobile phones, but unique to the CDN herein, they also retail removable memory cards pre-loaded with high-demand web content. When card capacity reaches the extreme, they will retail memory cards pre-loaded with the entire static Web.
- Mobile Phones provide the processing power to search and deliver content from the removable media; and to connect to the Web where available.

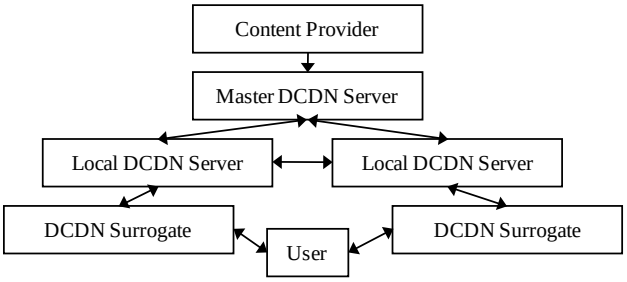

Fig. 1. DCDN Architecture - users are surrogate servers

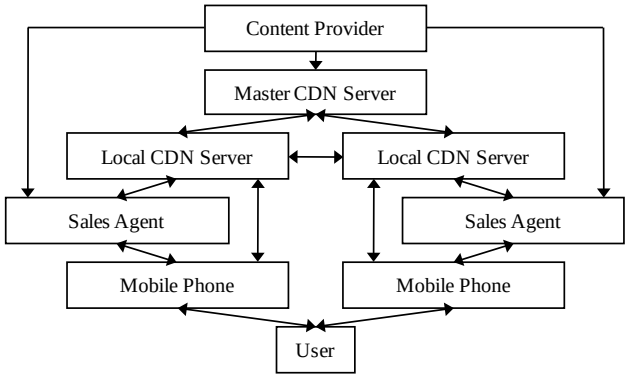

Fig. 2. CDN Architecture with sales agents and end users

This scheme leverages services provided by existing content delivery networks. It also leverages services provided by existing phone distribution networks. And it takes advantage of the storage and processing capacity of today's mobile phones. As such, it is realizable today.

The scheme offers two methods of content delivery: online and offline. The static and popular parts of the Web are available to both while the dynamic content is only available online.

With the static pages preloaded this scheme reduces Internet (and mobile) bandwidth, reduces server load, but most importantly it makes content available when the user is in a remote location – essential, for example, for the delivery of medical information to remote communities. It also reduces server response times from sites delivering static content as the content is stored locally and there is no round trip to a remote server.

Problematically it requires close collaboration between mobile sales outlets, and content delivery networks as protocols must be installed.

It also requires content providers to choose to take advantage of the new network. But there are no significant political hurdles to using CDNs as many content providers already use Akamai or Google to deliver content. Content providers also already accept large scale Internet Service Provider (ISP) proxying and long duration user-based caching, and are unlikely to object that what is analogous to a massively distributed cache. As user-based cached copies often stay on user's computer for more than a year [2] the thought of being able to provide more recent copies of their data on removable media could be appealing to a content provider. Some CDN providers

such as Akamai are already working with mobile companies such as Ericsson to provide a better service to mobile users[2] and it is reasonable to expect these organizations to take full advantage of the new CDN.

There could be a perceived freshness of content issue in the mind of the user. Prior work has shown that more than 25% of all web sites are updated less frequently than once per year (The median web page age is around 100 days) [4]. Mobile phone users also tend to change phone every 18-24 months as most people lock themselves into contracts of this duration, upgrading their phone with each new contract. In the proposed scheme the static Web will be bundled with each phone and concerned users will be able to purchase updates from retail outlets.

IV. FEASIBILITY

This section examines the feasibility of the proposal from the perspective of Web size and phone storage capacity.

A. Size of the Web

In 2008 Google claimed that their web crawler crawled over one trillion unique URLs [5]. However many of these pages are automatically generated, are not useful to users (such as spam) or are duplicates. Only a small proportion of the Web is useful, content-bearing pages.

As an example, Google claims to have the most comprehensive index of any search engine, but the history of number of documents they index suggests it is only a small proportion of those crawled. Google is known to have had 1 billion pages in 2000 [5], and 8 billion in 2004[3]. Recent estimates suggest Google's index currently contains about 30 billion documents[4], or about 3% of the URLs crawled. Despite this and the effort Google invests in spam and duplicate detection, users still observe both in results lists.

In 2010 Ramachandra [6] estimated the size of the average (uncompressed) web page by sampling a billions web pages. At that time the average web document was 380KB with about 37KB of that being the HTML file.

Assuming the search engine indexes 30 billion pages at 380KB each, the total searchable Web is 10,617TB. After compression (assuming a modest ratio of 0.2), the total size is at least 2,123TB. Assuming the index is 10% of the HTML file size the index file is at least 103TB.

B. Memory Card Capacity

This section discusses SD memory cards because they are the leader and de facto standard in flash memory [7]. Early data (1999-2003) is sourced from the SD 3.00 announcement [7], mid data (2005-2009) is sourced from the Panasonic SDXC cards development road map [8], recent data (2010) is from Sandisk's web site[5] and this year's data (2011) data is from the Macworld[6] article on Lexar.

Growth in SD memory card capacity between 1999 and 2011 is shown in Fig.3. Memory card capacity grew exponentially doubling almost every year between 1999 and 2009. In 1999 the capacity jumped from 8MB to 64MB not only doubling but octupling; it then slowed down but continued to double every two years. Current capacity (2011) is 128GB.

The current SD 3.0 specification for SDXC flash memory cards includes 2TB cards [8]. Assuming developments will continue at the current rate (doubling every two years), by 2020 a 2TB card will be on the market. For comparison, this is larger than the hard-disk capacity of many desktop PCs today.

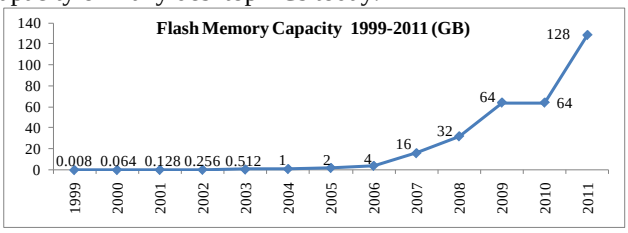

Fig. 3. Flash memory card capacity growth 1999-2011

Today's smart-phones already contain powerful CPUs similar to desktop CPUs of only a few years ago. The HTC EVO 3D, for example has a 1.2 GHz Dual-core processor and 1GB internal RAM and a microSD storage slot[7]. It is reasonable to believe that the core count will follow Moore's law.

C. Storing Web Content on Mobile Phones

This section examines how many web documents a mobile phone with 2TB of storage can carry.

A 2TB card can hold at least 5 million Web documents averaging 380KB in size. With a compression ratio of 0.2, by 2020 a mobile phone will be able to carry at least 25 million documents.

A rule of thumb for web page design is that a page should contain 250-300 words or around 1000 words for complex topics. The average adult reading rate is about 250 words per minute[8], or about one web page per minute or 480 pages per (8 hour) day. A reader reading 8 hours each day for 365 days per year would take nearly 30 years to read 5 million uncompressed documents on the phone.

V. DISCUSSION

A lifetime of reading can already be carried in the pocket, but what is it that we should read? There are several strategies for choosing these 25 million documents. One way is simply choosing the top 25 million most frequently requested pages. Existing research has

[3] See their homepage on the Wayback Machine: http://www.archive.org
[4] http://www.worldwidewebsize.com/
[5] www.sandisk.com
[6] www.macworld.com/article/156815/2011/01/lexar_128gbsdxc.html
[7] www.phonegg.com/Top/Fastest-Processor-Cell-Phones.html
[8] en.wikipedia.org/wiki/Words_per_minute

shown that the popularity of requested and transferred pages follows a Zipf-like distribution and the popularity of Websites or requests to servers follows a power law distribution (sometimes referred as "80/20 rule")[9]. Therefore, we suggest using the top most requested documents from the search engines to produce the bundle of web documents to be preloaded to mobile phones.

Personalized bundles containing relevant information to each mobile user can also be produced taking into account preferences of the user gathered directly from the user or indirectly from online social network information of the user (e.g. information available on FaceBook, Twitter, LinkedIn etc). Recommendation techniques exploiting the social network relationships between mobile users can also be used to deliver relevant content to mobile users.

An alternative to the popular but general bundles, and very personalized ones, is to suggest specialization based on either language, geographic "home" of the user, or topic (such as medicine, technical, etc).

Two sample specialized topical collections are Medline and the Wikipedia. A snapshot of the text of the English Wikipedia is included in the ClueWeb09 dataset[9] crawled in January and February 2009. It is about 6 million documents totaling (compressed) about 47.7GB. Our search engine's index of this is about 5.3GB and the size of the search engine software itself is negligible (less than 2MB). This totals 53GB, easily storable on today's 128GB cards with room to spare for (thumbnail) images.

Medline 2011 contains about 19,569,568 bibliographic references to, and abstracts of, academic papers in medicine and biology totaling 83.4GB uncompressed (11.1GB compressed)[10]. An index at 10% of the text size would be 8GB. This would fit on today's cards. In fact, Medline and the Wikipedia could already be stored on the same card!

Google web search indexes about 30 billion pages estimated (in the previous section) to be about 2,100TB of data. This will not fit on the 2TB phone card of 2020. The index file itself is estimated to be about 106TB. This will also not fit in the phone of 2020.

To sum up, it will be infeasible to fit the entire web on a mobile phone for the foreseeable future. However, important subsets such as the Wikipedia and Medline already fit, could be bundled, and updates could be pushed to the devices via SD card upgrades. This would make this important information available in remote locations where there is no Internet connectivity. It can also reduce access costs to knowledge on the web as there will be no Internet access fees when searching locally on the phone. With removable media, many important collections could be made available on separate cards thus bringing, for example, medical information to extremely remote locations on a device that is small, light-weight, portable, and inexpensive. This also provides mobile operators the opportunity to offer richer mobile services.

[9] http://boston.lti.cs.cmu.edu/clueweb09/wiki

[10] www.nlm.nih.gov/bsd/licensee/2011_stats/baseline_med_filecount.html

VI. CONCLUSIONS

Mobile phones are a far more important communication technology than the land-line Internet, especially for developing and under-developed countries without infrastructure. The rapid growth of storage capacity has already made it possible to bundle essential document collections such as the Wikipedia and Medline along with a search engine and the index. Doing so would provide information access anywhere anytime without causing Internet traffic or incurring access fees. We have proposed a new Web content delivery scheme. We have also proposed the method for content updates but acknowledge that it may not be necessary given the rate at which users replace their phones. This paper has identified a new direction for providing more reachable mobile web services.

Our further work will include researching which parts of the Web should be included, and whether these are domain specific (such as health, construction, and gaming) or general purpose. Techniques for generating personalized bundles that can be preloaded onto mobiles phones will also be investigated.

REFERENCES

[1] M. Hofmann, L. R. Beaumont, *Content Networking: Architecture, Protocols, and Practice*: Morgan Kaufmann, 2005.

[2] J. Charzinski, "Traffic Properties, Client Side Cachability and CDN Usage of Popular Web Sites," in *Measurement, Modelling, and Evaluation of Computing Systems and Dependability and Fault Tolerance*. vol. 5987, B. Müller-Clostermann, K. Echtle, E. Rathgeb, Eds.: Springer, 2010, pp. 136-150.

[3] J. P. Mulerikkal, I. Khalil, "An Architecture for Distributed Content Delivery Network," *ICON 2007*, pp. 359 - 364.

[4] B. E. Brewington, G. Cybenko, "How dynamic is the Web?," *Computer Networks*, vol. 33, pp. 257-276, 2000.

[5] J. Alpert, N. Hajaj, "We knew the web was big," http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html, accessed on: 04-Mar 2011.

[6] S. Ramachandra, "Web metrics: Size and number of resources,"http://code.google.com/speed/articles/web-metrics.html, accessed on: 13-Mar 2011.

[7] Y. Lin, "New Generation SD 3.00 Overview," in *Flash Memory Summit* Santa Clara, CA, USA, 2009.

[8] R. Lai, "Panasonic SDXC cards roadmap and Lumix camera lineup at CES 2010," http://www.engadget.com/2010/01/08/panasonic-sdxc-cards-roadmap-and-lumix-camera-lineup-at-ces-2010/, accessed on: 15-Mar 2011.

[9] Jiming Liu, Shiwu Zhang, J. Yang, "Characterizing Web Usage Regularities with Information Foraging Agents," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, pp. 1-19, 2004.