A Focused Mobile Web Search Engine Using a Topic-Specific Knowledge Base

Wen-Chen Hu
Department of Computer Science
University of North Dakota
Grand Forks, ND 58202-9015
wenchen@cs.und.edu

Sampath Bemgal
Rockwell Collins, Inc.
Cedar Rapids, IA 52498
sjbemga9@rockwellcollins.com

Lixin Fu
Department of Computer Science
University of North Carolina, Greensboro
Greensboro, NC 27402-6170
lfu@uncg.edu

Hung-Jen Yang
Department of Industrial Technology Education
National Kaohsiung Normal University
Kaohsiung, Taiwan 802
hjyang@nknucc.nknu.edu.tw

Abstract

Mobile knowledge seekers often need to access information on the World Wide Web during a meeting or on road, while away from their desktop computers. A common approach is to use Internet-enabled mobile handheld devices such as personal digital assistants or smart cellular phones to search the Internet. However, finding the appropriate queries for web search engines is never an easy task and the search results may or may not be relevant to what the users are really looking for. Also, constraints inherent in handheld devices such as slow communication, low storage capacity, and awkward input methods must be considered. Due to the above factors, information discovery using handheld devices becomes impractical and inconvenient. This paper investigates a new, innovative focused search method for handheld devices by using a topic-specific knowledge base.

1. Introduction

Using Internet-enabled mobile handheld devices to accessing the World Wide Web is a promising addition to the Web and traditional e-commerce. Mobile handheld devices provide convenience and portable access to the huge information on the Internet for mobile users from anywhere and at anytime. However, finding the appropriate queries for web search engines is never an easy task and the search results may or may not be relevant to what the users are really looking for. Also, constraints inherent in handheld devices such as slow communication, low storage capacity, and awkward input methods must be considered. Due to the above factors, information discovery by using handheld devices becomes impractical and inconvenient.

This research investigates a new, innovative focused search method for handheld devices by using a topic-specific knowledge base instead of the entire Web. The system performs the following three tasks:

1. Autonomous crawlers are sent to the (mobile) Internet to collect topic-specific data. Two kinds of data storage, a knowledge base on a server and persistent storage on a device, will be updated accordingly to reflect the latest content.
2. Users are able to search the data stored in the persistent storage on the devices off-line, to receive the results instantly.
3. If they do not like the results from the local searches, they can then search the entire knowledge base on the server.

Also, query formulation assistance for query composition will be provided to effectively find what the users want and proper user interfaces will be developed to show the search results. This approach is proved to greatly improve the mobile search results and speed. Other related issues such as client- and server-side handheld computing, and mobile commerce systems and transactions will also be discussed in this research.

2. Related Research

This section first introduces a mobile commerce system and then illustrates how it is used to perform a mobile transaction.

2.1 A Mobile Commerce System Structure

A mobile commerce system is inherently interdisciplinary and could be implemented in various ways. Figure 2 shows the structure of a mobile commerce system and a typical example of such a system (Hu, Lee, & Yeh, 2004). The system structure includes six components: (i) mobile commerce applications, (ii) mobile handheld devices, (iii) mobile middleware, (iv) wireless networks, (v) wired networks, and (vi) host computers.

1. *Mobile commerce applications:* Electronic commerce applications are numerous, including auctions, banking, marketplaces and exchanges, news, recruiting, and retailing, to name but a few. Mobile commerce applications not only cover the electronic commerce applications,

but also include new applications, which can be performed at any time and from anywhere by using mobile computing technology, for example, mobile inventory tracking.

2. *Mobile handheld devices:* An Internet-enabled mobile handheld device is a small general-purpose, programmable, battery-powered computer that is capable of handling the front end of mobile commerce applications and can be operated comfortably while being held in one hand. A mobile handheld device includes six major components: (i) a mobile operating system, (ii) a mobile central processor unit, (iii) a microbrowser, (iv) input/output devices, (v) a memory, and (vi) batteries (Hu, Yeh, Chu, & Lee, 2005).
3. *Mobile middleware:* The major task of mobile middleware is to seamlessly and transparently map Internet contents to mobile stations that support a wide variety of operating systems, markup languages, microbrowsers, and protocols. WAP and i-mode are the two major kinds of mobile middleware. According to an article in (Eurotechnology, n.d.), 60 percent of the world's wireless Internet users use i-mode (NTT-DoCoMo, n.d.), 39 percent use WAP (Open Mobile Alliance, n.d.), and 1 percent use Palm middleware. Table 1 compares i-mode and WAP.

	WAP	**i-mode**
Developer	WAP Forum	NTT DoCoMo
Function	A protocol	A complete mobile Internet service
Host Language	WML	CHTML
Major Technology	WAP Gateway	TCP/IP modifications
Key Features	Widely adopted and flexible	Highest number of users and easy to use

Table 1: Comparisons between the Two Major Types of Mobile Middleware.

4. *Wireless and wired networks:* Wireless communication capability supports mobility for end users in mobile commerce systems. Wireless LAN, MAN, and WAN are the major components used to provide radio communication channels so that mobile service is possible.
5. *Host computers:* A user request such as database access or updating is actually processed at a host computer, which contains three major kinds of software: (i) Web servers, (ii) database servers, and (iii) application programs and support software.

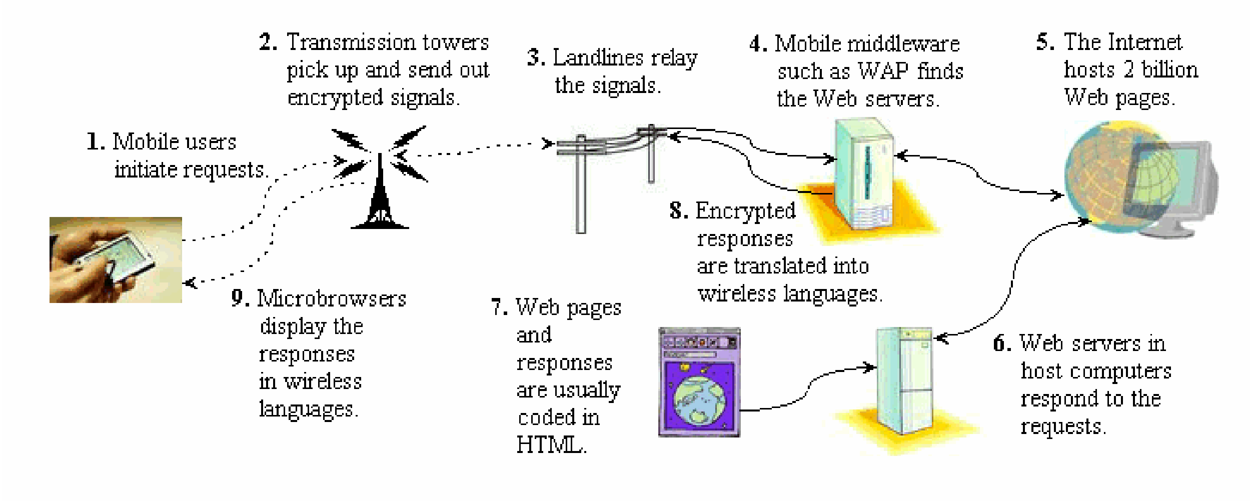

2. Transmission towers pick up and send out encrypted signals.
3. Landlines relay the signals.
4. Mobile middleware such as WAP finds the Web servers.
5. The Internet hosts 2 billion Web pages.
1. Mobile users initiate requests.
8. Encrypted responses are translated into wireless languages.
9. Microbrowsers display the responses in wireless languages.
7. Web pages and responses are usually coded in HTML.
6. Web servers in host computers respond to the requests.

Figure 1: A Flowchart of a User Request Processed in a Mobile Commerce System.

2.2 An Example of Mobile Commerce Transaction Processing

To explain how the mobile commerce components work together, Figure 1 shows a flowchart of how a user request is processed by the components in a mobile commerce system, along with brief descriptions of how each component processes the request.

1. *Mobile commerce applications:* A content provider implements an application by providing two sets of programs: client-side programs, such as user interfaces on microbrowsers, and server-side programs, such as database access and updating.
2. *Mobile handheld devices:* Handheld devices present user interfaces to the mobile end users, who specify their requests on the interfaces.
3. *Mobile middleware:* The major purpose of mobile middleware is to seamlessly and transparently map Internet contents to mobile stations that support a wide variety of operating systems, markup languages, microbrowsers, and protocols.
4. *Wireless networks:* User requests are delivered to either the closest wireless access point (in a wireless local area network environment) or a base station (in a cellular network environment).
5. *Wired networks:* This component is optional for a mobile commerce system. However, most computers (servers) usually reside on wired networks such as the Internet, so user requests are routed to these servers using transport and/or security mechanisms provided by wired networks.
6. *Host computers:* Host computers process and store all the information needed for mobile commerce applications, and most application programs can be found here. They include three major components: Web servers, database servers, and application programs and support software.

3. The Proposed Focused Search Engine

A mobile web focused search system is proposed in this research (Bemgal, 2006). Figure 2 shows the system structure of a typical search engine (Hu, Yang, Yeh, & Lee, 2004). Search engines traditionally consist of three major components: (i) the crawler, (ii) the indexing software, and (iii) the search and ranking software:

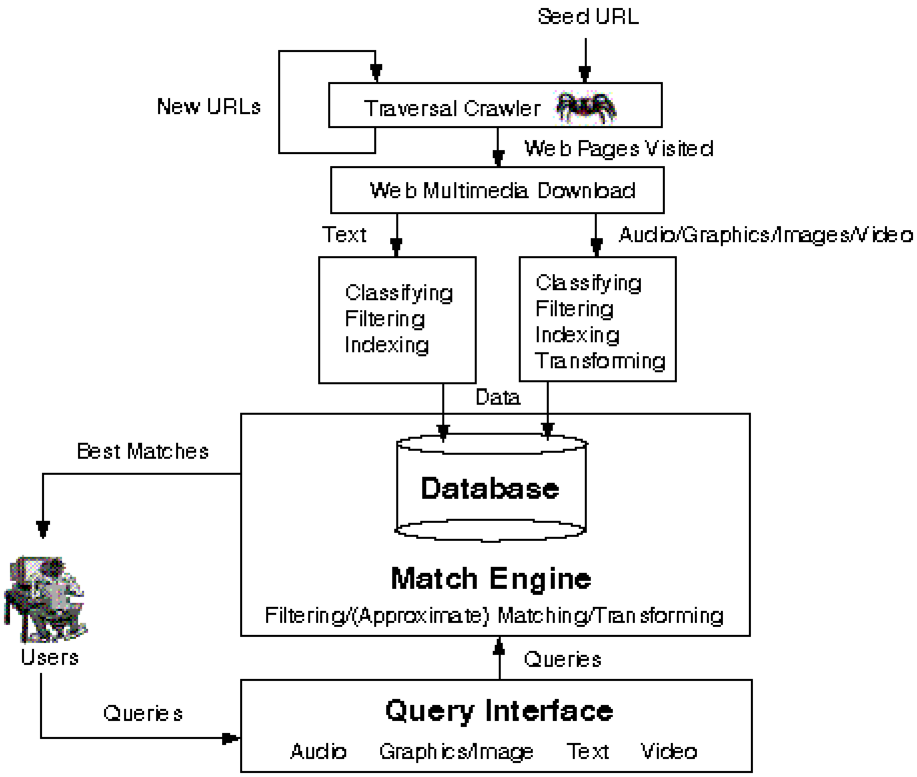


Figure 2: System Structure of Web Search Engine.

- A crawler is a program that automatically scans various web sites and collects web documents from them. Crawlers follow the links on a site to find other relevant pages. Two search algorithms, breadth-first searches and depth-first searches, are widely used by crawlers to traverse the Web.
- Automatic indexing is the process of algorithmically examining information items to build a data structure that can be quickly searched. Traditional search engines utilize the following information, provided by HTML scripts, to locate the desired web pages: (i) content, (ii) descriptions, (iii) hyperlink, (iv) hyperlink text, (v) keywords, (vi) page title, (vii) text with a different font, and (viii) the first sentence.
- Query processing is the activity of analyzing a query and comparing it to indexes to find relevant items. A user enters a keyword or keywords, along with Boolean

> modifiers such as “and,” “or,” or “not,” into a search engine, which then scans indexed Web pages for the keywords. To determine in which order to display pages to the user, the engine uses an algorithm to rank pages that contain the keywords.

Details of the three components of the proposed focused web search engine are given next.

3.1 The Crawler

A crawler is also known as robot or agent. The crawlers are the programs that automatically explore the World Wide Web by retrieving a document and recursively retrieving some or all the documents that are referenced in it (Menczer, Pant, & Srinivasan, 2004). Crawlers follow the links on a site to find other relevant pages. There are generally two algorithms involved in crawlers: one is depth-first search and the other is breadth-first search. A breadth-first search is applied in this research. Figure 3 shows the crawler structure, where an “array” is used for the queue for storing the URLs. Initially, a seed URL, given by the system administrator, is stored in the queue, which forms the basis for the crawler to explore the Web. The crawler takes one URL at a time and retrieves the content of the selected URL. The content is then parsed, for example, the keywords are collected and saved. The URLs referred by this page are added to the end of the queue and the process is repeated for each URL. In this way, the information for the knowledge base is collected. The crawling stops when the queue is empty.

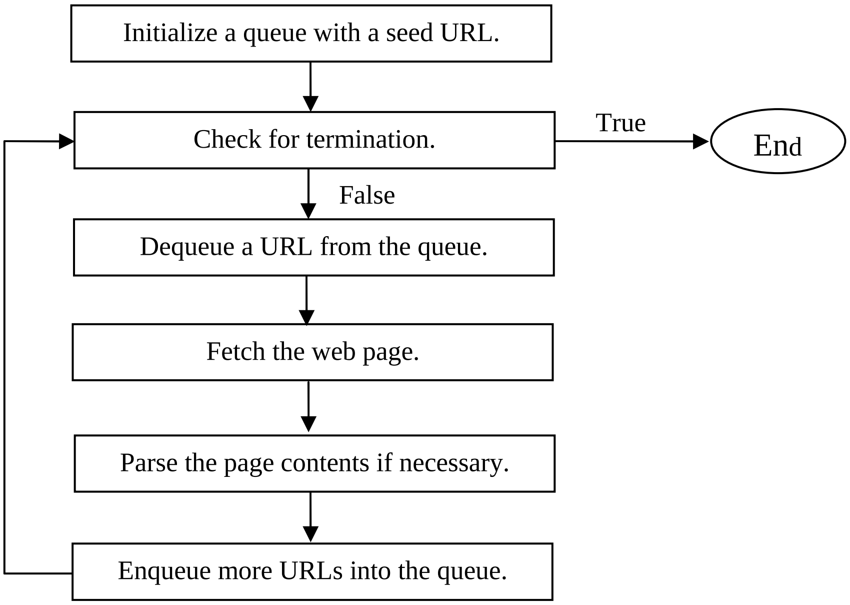

Initialize a queue with a seed URL.
Check for termination.
True
End
False
Dequeue a URL from the queue.
Fetch the web page.
Parse the page contents if necessary.
Enqueue more URLs into the queue.

Figure 3: A Crawling Procedure.

3.2 Indexing Using a Knowledge Base

A knowledge base used in this research is a repository storing web data for searching. The knowledge about a topic of interest can be represented as a set of thematically related URLs or as a set of queries (Aridor, et al, 2001). The first approach is applied in this system where a seed URL is given for the crawler to explore the Web initially. The knowledge base constructed is comprised of topic-related terms and URLs. This knowledge base could help users to retrieve the topic-related information as it is constructed with the related information only. The crawler is responsible for knowledge acquisition. In general, several crawlers are used to acquire data. The knowledge acquisition is considered as an evolutionary process. The knowledge base is evolved by iteratively collecting URLs from the Web starting with a seed URL. Once the data is collected, it is processed, indexed, and stored in the knowledge base. In general, there are many indexing strategies. This research creates a web-document index by mainly using URLs and hyperlinked text in a document. The indexed data is stored in three places:

- *A knowledge base on the server:* It includes a set of database tables, which store the retrieved URLs and URL text. Its size is about few megabytes limited by a student account.
- *A persistent storage in a device:* This footprint is actually implemented by using files in the server. It stores a small part of the data in the knowledge base and is limited to few hundreds of kilobytes.
- *A cache in a device:* It is simulated by creating a space of tens of kilobytes in the server. It contains a small part of the data in the footprint or knowledge base.

Though simulation is applied to the last two methods, it should not be a problem if the storage is really built on the devices.

3.3 Searching and Ranking

When a user submits a query, the search engine will go through the indexes to find the relevant web pages and display the search results on the screen. The order of the displayed results is based on some ranking methodology. This research proposes a query formulation assistance to facilitate mobile data search and entry, which is usually a problem for mobile users. It provides users a list of related terms by using the lexicon stored in the knowledge base. Other than entering the query keywords, the users can pick related terms to help the system perform better searches and reduce the amount of data entry. Once the query is formulated and submitted, the search engine performs the matching and displays the search results by the number of keyword matched. Various search filtering operations such as stopword deletion and space removal to name a few are implemented.

4. Experiments

This section gives a few screenshots demonstrating various information discoveries used in this research. There are, however, some factors need to be considered when implementing the entire system:

- The first is the number of URLs stored in the knowledge base. Due to limited capacity of the Oracle student account, only few URLs are stored, though a large number of URLs are generated. The page content is not stored, which could be useful.
- Second, there are several WAP browsers available for mobile application testing and implementation (WinWap Technologies, n.d.). One of them is "TTEmulator," which is used in this research. It is an online WAP emulator hosted at the http://www.tagtag.com/ (Inetis Ltd., n.d.).
- Finally, it is the mobile interface usability (Buchanan, et al, 2001). Mobile interface is usually small and inconvenient for users. This research proposes innovative mobile interface to make the entire system feasible and easy to use.

4.1 Experimental Results

Initially, a set of topics will be presented to users as a drop-down list and another option allows users to create their own topics of interest. Those topics are hosted by a knowledge base, which is stored in the backend server. A small part of the knowledge base is also stored in the persistent storage of a device. Users can then perform searching on the knowledge base or the footprint on the device. A query formulation assistance is also provided in this system. Users can refine his/her queries for better results. Since there is a small footprint of the knowledge base available on the device, searching can be performed off-line too. If the search results from the footprint are not of interest to the users, then the entire knowledge base can be searched. Users can also cache the search results and perform searching on those cached results later. The following two demonstrations will show the strength of this system.

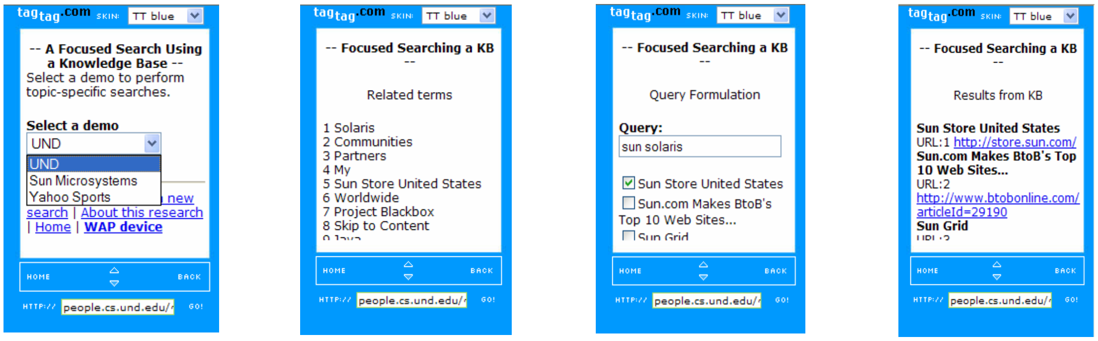
Figure 4 gives a demonstration of searching the web pages of Sun Microsystems at http://www.sun.com/ by using the knowledge base:

(a) Give the system entry page, which allows users to pick one of the three demonstrations: UND, Sun Microsystems, and Yahoo Sports.

(b) Display a list of relevant terms, which are the hyperlinked text.

(c) Show the query formulation interface, which allows users to provide additional information other than the query.

(d) List the search results from the knowledge base.



-- A Focused Search Using a Knowledge Base --
Select a demo to perform topic-specific searches.

Select a demo
UND
UND
Sun Microsystems
Yahoo Sports
new search | About this research | Home | WAP device

-- Focused Searching a KB --

Related terms

1 Solaris
2 Communities
3 Partners
4 My
5 Sun Store United States
6 Worldwide
7 Project Blackbox
8 Skip to Content
9 Java

-- Focused Searching a KB --

Query Formulation

Query:
sun solaris

☑ Sun Store United States
□ Sun.com Makes BtoB's Top 10 Web Sites...
□ Sun Grid

-- Focused Searching a KB --

Results from KB

Sun Store United States
URL:1 http://store.sun.com/
Sun.com Makes BtoB's Top 10 Web Sites...
URL:2 http://www.btobonline.com/articleId=29190
Sun Grid
URL:3

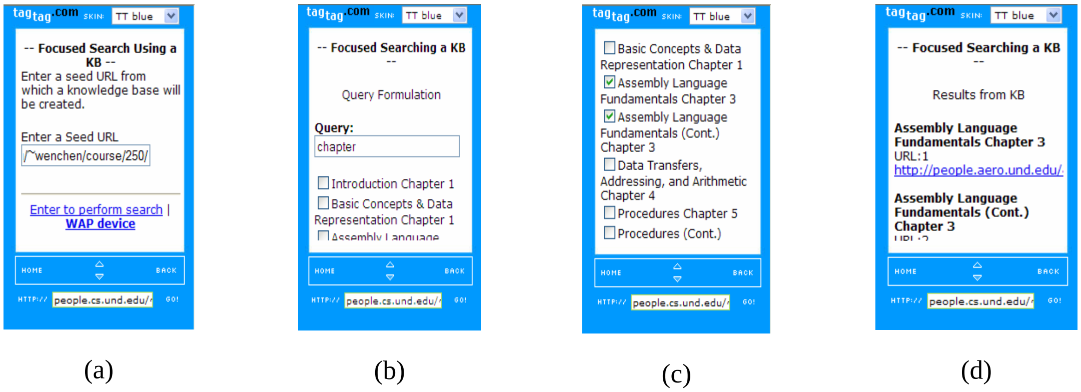
(a) (b) (c) (d)

Figure 4: A Demonstration of Searching “Sun Microsystems.”

The above demonstration shows the searches on the pre-set data. This system also allows users to dynamically collect and search data. Figure 5 gives another demonstration showing dynamic knowledge collection and searching:

(a) An interface allows users to enter a seed URL for knowledge collection.
(b) & (c) After the knowledge base is constructed, users can start performing searching by using the query formulation assistance.
(d) Show the search results from the knowledge base.

Users can also perform their searches in a disconnected mode by using the footprint or cache in the device.



(a)

-- Focused Search Using a KB --
Enter a seed URL from which a knowledge base will be created.

Enter a Seed URL
/~wenchen/course/250/

Enter to perform search | **WAP device**

people.cs.und.edu/

(b)

-- Focused Searching a KB --

Query Formulation

Query:
chapter

□ Introduction Chapter 1
□ Basic Concepts & Data Representation Chapter 1
□ Assembly Language

people.cs.und.edu/

(c)

□ Basic Concepts & Data Representation Chapter 1
☑ Assembly Language Fundamentals Chapter 3
☑ Assembly Language Fundamentals (Cont.) Chapter 3
□ Data Transfers, Addressing, and Arithmetic Chapter 4
□ Procedures Chapter 5
□ Procedures (Cont.)

people.cs.und.edu/

(d)

-- Focused Searching a KB --

Results from KB

Assembly Language Fundamentals Chapter 3
URL:1
http://people.aero.und.edu/

Assembly Language Fundamentals (Cont.) Chapter 3
URL:2

people.cs.und.edu/

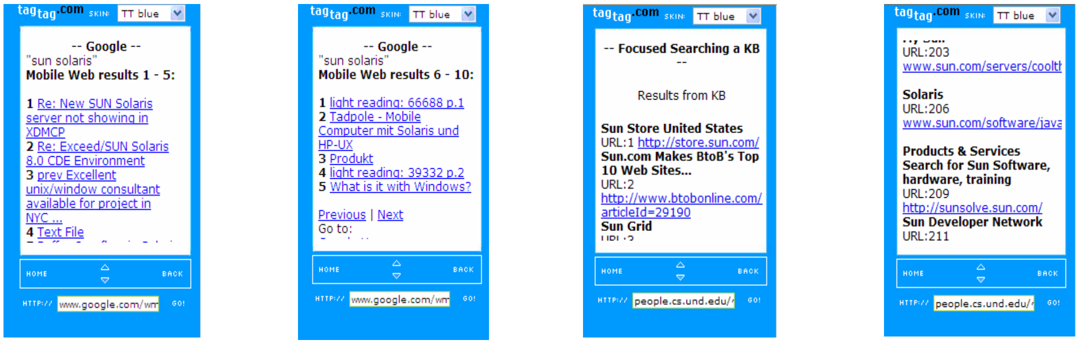
Figure 5: Another Demonstration Showing Dynamic Knowledge Collection and Searching.

4.2 Comparisons

From the previous experimental results, the proposed methods increase the search speed and relevance between the search results and query. All search results from our methods are related to the query either directly or indirectly. No irrelevant results are displayed. This sub-section compares our search results to the results from the Google WAP Search (Google, n.d.). Figure 6 gives the screendumps from both systems for the query “sun solaris”:

- Figures 4.a and 4.b show the top-10 search results from the Google WAP Search and
- Figures 4.c and 4.d show the top-6 search results from our system.

Google provides more search results, but some of the top results are not related to the query, which might take time from users to find the desired results. On the other hand, our search results are all related to the query.



-- Google --
"sun solaris"
Mobile Web results 1 - 5:

1 Re: New SUN Solaris server not showing in XDMCP
2 Re: Exceed/SUN Solaris 8.0 CDE Environment
3 prev Excellent unix/window consultant available for project in NYC ...
4 Text File

www.google.com/wm

(a)

-- Google --
"sun solaris"
Mobile Web results 6 - 10:

1 light reading: 66688 p.1
2 Tadpole - Mobile Computer mit Solaris und HP-UX
3 Produkt
4 light reading: 39332 p.2
5 What is it with Windows?

Previous | Next
Go to:

www.google.com/wm

(b)

-- Focused Searching a KB --

Results from KB

Sun Store United States
URL:1 http://store.sun.com/
Sun.com Makes BtoB's Top 10 Web Sites...
URL:2
http://www.btobonline.com/articleId=29190
Sun Grid

people.cs.und.edu/

(c)

URL:203
www.sun.com/servers/coolth

Solaris
URL:206
www.sun.com/software/java

Products & Services
Search for Sun Software, hardware, training
URL:209
http://sunsolve.sun.com/
Sun Developer Network
URL:211

people.cs.und.edu/

(d)

Figure 6: Search Results from the Proposed Methods and Google.

Table 1 summarizes the differences between our approach and Google's WAP Search. From the above experimental results, this approach tends to increase the relevance between the search results and query compared to commercial search engines. Though the search engine implemented may not vie with commercial search engines in scale, this research shows the effectiveness of the proposed methods and concepts. The other disadvantage of using Google's WAP Search or any other search engines is that some irrelevant results are displayed, which might distract users and thus waste their time.

	The Proposed Approach	**Google**
Result Relevancy	Results closely related to the query	Some results related to the query but some not
Number of Results	Less but all related to the query directly or indirectly	More but after the first few results, the relevance decreasing
Search Speed	Fast because of using query formulation assistance	Slow because of no query formulation
Options	Caching results and checking related URLs and terms	No advanced features available

Table 1: Comparisons between the Proposed Approach and Google's WAP Search.

5. Conclusions

This research proposes a new approach for mobile information retrieval. It applies a focused search to a topic-specific knowledge base, whose topics are selected by system administrators. The knowledge base is located at a server, which can be accessed using wireless communication. Users can select any one of the topics and start searching the knowledge base. A small part of the knowledge base can also be downloaded to the persistent storage in a device and search can be conducted locally in a disconnected mode. If the users do not like the results from local searches, they can search the entire knowledge base. The advantages and disadvantages of using this system are as follows:

- *Advantages:*
  - The users need not search the entire (mobile) Internet to find relevant information. Instead, they can select the topic of interest and perform a search only on the information related to the topic.
  - In the absence of wireless access, information can still be searched and retrieved since a small footprint is included in the device.
  - Navigation and browsing are greatly improved with the help of the proposed interfaces.
  - Query formulation assistance is provided to facilitate searching and date entry.
- *Disadvantages:*
  - Since server-side methodology is used in this research, there might be some unseen difficulties if the approach is implemented on devices, the client side.
  - Because of the small storage space used, the indexed topics and data are limited. For example, the information requested might not be stored at all.

Though only server-side handheld computing is used in this research, this approach can also be applied to client-side devices. There are several environments/languages used in client-side handheld computing such as BREW, J2ME, Palm OS, Symbian, and Windows Mobile. For example, the data can be stored in persistent storage of a device, such as Record management system of J2ME. The future works will include extending this approach to different types of handheld devices. Other future works are

- Implement and strengthen the system on a large scale by using both client-side and server-side handheld computing.
- Save more information in the knowledge base such as text of web pages and audio/video contents, so that mobile information retrieval can be improved tremendously.

With more people using handheld devices, mobile information discovery becomes crucial. While current devices may not posses enough memory or processing capabilities to apply the proposed methods, hopefully these devices will have enough features to handle the proposed methods and many other complex tasks in the near future.

References

Aridor, Y., Carmel, D., Lempel, R., Maarek, Y. S., & Soffer, A. (2001). Knowledge encapsulation for focused search from pervasive devices. In *Proceedings of the WWW10 Conference*, pp. 754-764.

Bemgal, J. (2006). *A Focused Search Using a Topic-Specific Knowledge Base for Web-Enabled Mobile Handheld Devices*. Unpublished master’s thesis, University of North Dakota.

Buchanan, G., Farrant, S., Jones, M., Thimbleby, H., Marsden, G., & Pazzani, M. (2001). Improving mobile Internet usability. In *Proceedings of the 10th International WWW Conference*, Hong-Kong.

Eurotechnology. (n.d.). *Frequently Asked Questions about NTT-DoCoMo's i-mode*. Retrieved October 9, 2006, from the World Wide Web: http://www.eurotechnology.com /imode/faq.html

Google. (n.d.) *Google's WAP Search*. Retrieved March 23, 2007, from the World Wide Web: http://www.google.com/wml/

Hu, W.-C., Lee, C.-w., & Yeh, J.-h. (2004). Mobile commerce systems. In Shi Nansi, editor, *Mobile Commerce Applications*, pages 1-23, Idea Group Publishing.

Hu, W.-C., Yang, H.-J., Yeh, J.-h., & Lee, C.-w. (2004). World Wide Web search technologies. In Mehdi Khosrow-Pour, editor, *Encyclopedia of Information Science and Technology*, I(V), pages 3111-3117, IRM Press.

Hu, W.-C., Yeh, J.-h., Chu, H.-J., & Lee, C.-w. (2005). Internet-enabled mobile handheld devices for mobile commerce. *Contemporary Management Research*, Vol. 1(1), pages 13-34.

Inetis Ltd. (n.d.). *TT Emulator*. Retrieved September 18, 2006, from the World Wide Web: http://www.inetis.com/index.php?module=ttemulator

Menczer, F., Pant, G., & Srinivasan, P. (2004). Crawling the Web. In M. Levene and A. Poulovassilis, editors, *Web Dynamics: Adapting to Change in Content, Size, Topology and Use*, Springer-Verlag.

NTT-DoCoMo. (n.d.). *i-mode*. Retrieved October 2, 2006, from the World Wide Web: http://www.nttdocomo. com/

Open Mobile Alliance, (n.d.). *WAP (Wireless Application Protocol)*. Retrieved September 14, 2006, from the World Wide Web: http://www.openmobilealliance.org/tech/affiliates/wap/wapindex.html

WinWap Technologies. (n.d.). *WinWAP Browsers*. Retrieved September 14, 2006, from the World Wide Web: http://www.winwap.com/products_2.php