Information Gathering and Presenting (IGP) for Mobile Shopping

By Kai ZHAO,* Hongwei QI * and Min-Yu HSUEH*

ABSTRACT Using the mobile environment (as opposed to the PC environment) as the primary channel for online shopping has not been attempted. The reasons for this are many, including the "PC" nature of the whole online shopping infrastructure and the technology limitations of the mobile environment. Nevertheless, in China, the huge number of mobile users and the large gap between the number of mobile and Internet users (340 million versus 100 million) give good market incentives to making mobile shopping a primary method for online shopping. In this paper we introduce the Information Gathering and Presenting (IGP) technology and its application in mobile shopping. The IGP technology collects and organizes relevant information, presents the precise information that the user is seeking, and helps the user avoid errors in his information search interactions with the system. These capabilities of IGP make it very suitable for information searches and interaction tasks in the mobile environment, where the interaction time is typically short, and the ability of the network to deliver large amounts of data is more constrained than in the PC-broadband case.

KEYWORDS **Mobile shopping, IGP (Information Gathering and Presenting), Information retrieval, Fault-tolerant search**

1. INTRODUCTION

The Information Gathering and Presenting (IGP) technology is being developed to produce precise and concise answers to searches in the Internet and other large information bodies. Today's popular search engines, such as Google (www.google.com), can return large amounts of relevant information to a query and leave it to the user to browse through such information. In contrast, IGP is designed to return a small number of much more accurate answers to a user's query.

This capability of IGP makes it ideal for information retrieval applications used in the mobile environment. The interaction time here is short and the interaction modes are not confined to the hand-eye, browse-and-click-links actions commonly used in the PC and Web environment. (Although not central to the thesis of this paper, it should be noted that IGP's precision and conciseness qualities may be applied to the PC-Web environment to streamline the user's information search efforts.)

In this paper, we introduce the key technical aspects of IGP: Information gathering and organization by context, information presentation, and fault-tolerant handling of user input. "Information gathering and organization by context" utilizes the knowledge that may be inferred from the context surrounding a piece of information to help categorize that information. "Information presentation" can pick out answers to a user's query from such categorized information to make a more precise and concise presentation than can today's search engines. "Fault-tolerant user input handling" further minimizes occurrences of spurious results caused by erroneous user input.

As an early validation of the IGP technology, it is being applied to mobile shopping of books in China. Online shopping has seen a great rise in popularity in recent years throughout the world. However, use of mobile devices in online shopping has been very limited. The premier online shopping company, Amazon.com (www.amazon.com), scaled back its mobile shopping platform "Amazon Anywhere" in the U.S. in 2002 due to lack of customer interest.

The early implementations of mobile shopping, such as "Amazon Anywhere" faced a number of difficulties ranging from technical to cultural and business. The "cultural" and business difficulties included the U.S. population's preference for using mobile phones almost exclusively for voice communication, and the extra cost of data communication that

*NEC Laboratories China

U.S. mobile operators charged.

In other countries, such as Japan, such hurdles are not as deeply rooted in the first place. In fact, in China, use of mobile devices to access small amounts of information and send short messages is the preferred mode of (data) communication due to its preferential economics over uses of PC, Web, and voice communication. Today, China has over 340 million mobile users and a comparatively small 100 million Internet users. The major online book shopping sites in China have completed an estimated two and a half million transactions (purchases) during 2004. Our mobile shopping implementation is aimed at popularizing this new mode of online (over air) shopping among the sizable 340 million Chinese mobile users.

2. THE INFORMATION GATHERING AND PRESENTING (IGP) TECHNOLOGY

The Information Gathering and Presenting (IGP) technology aims to deliver precise and concise answers to searches in large information bodies, including the Internet. Today's popular Internet search engines can return large amounts of relevant information to a search query, but the returned information often does not answer immediately the question that the user is asking.

For example, if it is required to find out how to travel from San Francisco International Airport to the city of Berkeley by searching on the Web, you could type the query "San Francisco International Airport to Berkeley" into any of the major U.S. search sites, such as Google, Yahoo (www.yahoo.com), or MSN (www.msn.com). Then you would have to pore through the search results and view the linked pages to discover for yourself about the low-cost (US$5.5) high-speed train service called BART (The Bay Area Rapid Transit, www.bart.gov).

In the following subsection, the lack of precision and compactness of the output of today's search engines is illustrated with an example. The example serves as a guide in the subsequent descriptions of the IGP technology.

2.1 Some Issues with Today's Search Engines

Table I shows the search results from Google of three similar queries for travel directions from the San Francisco International Airport to Berkeley. The only differences between the three queries are the omission of the word "International" from the second query and the use of San Francisco International Airport's airport code "SFO" in the third query. The search results, however, are quite different and none of the topmost results can help the user discover the BART service, unless he can read or infer from the Chinese from the "SFO" search. The same set of queries submitted to the other major U.S. search engines produced similar results. In fact, in all cases, the topmost return of the search with "SFO" contained the BART link; it appears that most Web page authors who know the way to Berkeley also speak airport jargons.

From this example and others, some of the key lacking areas of today's search engines are summarized below:

- Lack of knowledge or not taking advantage of knowledge offered by synonyms or similar concepts; for example, "SFO" and the "San Francisco International Airport" refer to the same

Table I The top five results from three searches for travel information from the San Francisco International Airport to Berkeley. Bold letters indicate Web pages with proper links to the BART rapid transit system.

Results	San Francisco International Airport to Berkeley	San Francisco Airport to Berkeley	SFO to Berkeley
1	UC Berkeley Public-Safety page with outdated info on BART ("no airport station").	UC Berkeley Public-Safety page with outdated info on BART ("no airport station").	**UC Berkeley student organization's travel directions from the airport. The page is in Chinese.**
2	Lonely Planet's San Francisco Guide.	**Blogon2004 conference travel information.**	**UC Berkeley Thai Student Organization travel directions.**
3	**J. Cletheroe's (a private person's) U.S. vacation hints.**	San Francisco Airport Limousine service to Berkeley.	**Concur company's training class travel directions.**
4	**Directions to the Lawrence Berkeley Lab.**	Driving directions to the Rose Garden Inn.	Elite Limousine service to Berkeley
5	Fodor's Berkeley Guide.	**Directions to the Lawrence Berkeley Laboratory.**	An individual's Web page under web.mit.edu.

entity;

- Lack of mechanisms for understanding users' intentions; Google, for example, ignores the preposition "to" in the query "San Francisco Airport to Berkeley," thereby not taking advantage of the user's hint that he is looking for travel directions. If this user intention was understood, the output could be a table of main transportation means between the airport and Berkeley;
- Not considering the freshness of information. This contributes to offering outdated (pre June 2003) information in two of the top returns in the example. In this case, the outdated information would make the user take an unnecessary bus ride to another BART station 12 kilometers north of the San Francisco International Airport.

The above observations are not surprising given that today's search engines are built to find relevant Web pages containing those keywords appearing in the query[1]. IGP's goals are different. IGP wants to understand the user's intention as much as possible, and give the user precise and concise answers.

2.2 The IGP Architecture

Figure 1 shows the conceptual diagram of the IGP architecture for handling text information from the Web.

The lower part of the architecture is responsible for information gathering. The Information Gatherer (IG) crawls the Web to discover useful information and information categories. Information categories

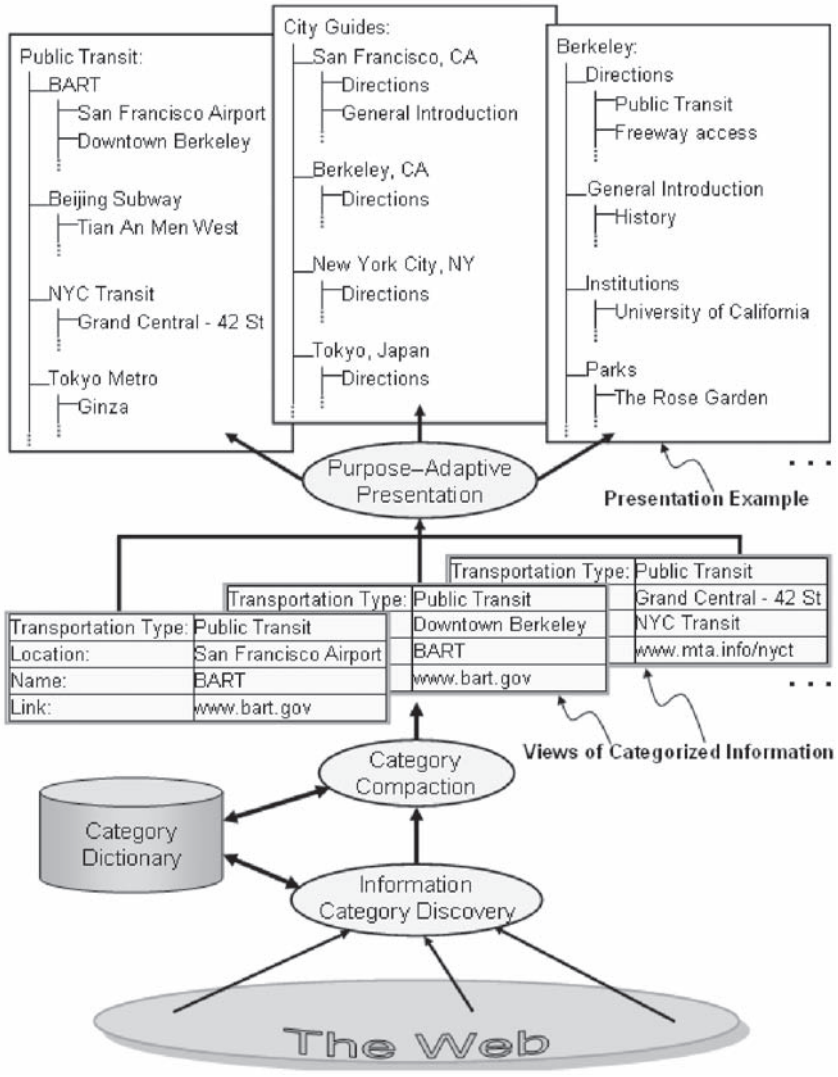

Fig. 1 A conceptual diagram of the IGP architecture.

are self-forming and updating, starting from seed categories in the Category Dictionary (CD). (The newly formed and updated information categories are deposited in the CD to become candidate seeds for future use.)

To start, the CD contains a number of seed categories derived from "common" knowledge. For example, a public transit system consists of lines and stations. IG does not require such seed categories to be "complete" or "defining" the world, as some of the work in Semantic Web attempt to accomplish[2]. Instead, IG will learn and update new categories as it finds relevant information. For example, IG may discover that public transit systems also offer schedule planners on their Web pages.

Another important task of IG is to compact categories by either merging them or linking them under appropriate contexts. For example, the name "San Francisco International Airport" may be first learned by IG under the "public transit station" category and the name "SFO" under the "airport code" category. Under the transportation context, the two categories are related and therefore linked and their constituent names, e.g., "San Francisco International Airport" and "SFO," compared and equated with the aid of other supporting information and categories.

The upper part of the architecture is responsible for presenting the information categorized information in a precise and concise manner to the user. At the present, IGP does not attempt to interpret user's queries in its full natural language form. Instead, IGP uses a combination of user's keywords input and selection from a response menu that IGP creates based on the keywords to determine the user's query intent.

As Fig.1 demonstrates, the categorized information may be presented in any suitable combination in replying to a user's query.

3. IGP FOR MOBILE SHOPPING

While creating a complete IGP technology is an ambitious endeavor and is still underway, our initial research results can already be applied to building a mobile book shopping flow. Online book shopping is one of the better understood and comparatively straightforward e-commerce processes.

3.1 Gathering and Organizing Information

Information categories for books are generally well-understood. They include the book title, the author, the publisher, the date of publication, etc. In fact, online bookstores show such information in well-organized Web pages for individual books.

In applying IGP for mobile book shopping, book information categories from our collaborating online bookstores are used as seeds in the category dictionary, or CD. With these seeds, IGP infers from the Web other possible categories for each book title. For example, Web pages about the American author Laura Wilder's popular children's book "Little House on the Prairie" include a large number of occurrences of such phrases as "frontier girl," "frontier woman," "American west," and dates in the 1800's. From these and information on the Web about other books, IGP infers additional book information categories "Story Period," "Story Location," and "Story Major Characters." The category "Story Period," for example, is inferred from the dates in the various Web pages that discuss the book's story. "Story Location" as a category is inferred from phrases "frontier" and "American west." The words "girl" and "woman" let IGP infer the category "Story Major Characters" in the context of a (fiction) book.

The purpose of such an extensive categorization of information is to gather all relevant information that can help increase the comprehensiveness and precision of the search result. For example, if a reader is looking for books on "stories of frontiers of the American west," IGP will find books with such words in their titles as well as books such as "Little House on the Prairie" because their category information contain matching words. In contrast, to obtain similar results, online bookstores have been including manually created book reviews as part of the search material.

However, the comprehensiveness achieved with the information gathering operation can easily produce an overwhelming amount of search results. The task of the information presenting part of IGP is to organize the search results in a concise manner.

3.2 Clustering Based Information Presenting

Category trees such as those in the upper part of Fig. 1 can present large amounts of information concisely. It is the main form of results presentation employed by IGP at the present. Thus, IGP's presentation task involves organizing the large amounts of search results into presentation categories and selecting the most relevant categories and category contents to show to the user.

3.2.1 Related Work

Before the advent of Web search engines, some researches focused on querying a pre-collection of documents. A static clustering of the entire collection

is made in advance. When a query arrives, it is matched to the cluster centroids and the top-ranked clusters are chosen to present to users[3]. However, this method is not suitable for the Web environment, because the collection of information for a Web search engine is usually too large and fluid[4]. Moreover, Reference [3] mentioned some experiments showing that this kind of clustering search does not outperform non-clustering searches except on some small collections.

Hearst and Pedersen changed the above static clustering method to a dynamic clustering method[5], that is, clustering on the retrieved documents of the query, but not on the whole pre-collection of documents. Since different documents are retrieved for different queries, the method is called dynamic (or online). The experiments showed significant improvements over similarity search ranking alone. Most of the later work on clustering searches follows this kind of clustering.

Since dynamic clustering is quite time consuming, fast algorithms have been developed. Zamir and Etzioni proposed the STC (Suffix Tree Clustering) algorithm [4,6], which was time linear in the size of the document set and reflected the sequence of words in the phrases. The cluster interface was called Grouper (it has ceased operation now), which was based on the HuskySearch meta-search engine that used traditional ranked-list presentation. The logs of the search engines showed that, users followed only 1.0 document on average in HuskySearch, compared with 1.4 document in Grouper. In addition, users spent less time browsing the result in Grouper than in HuskySearch when they followed more than three documents. These experiments under real environment conditions demonstrate the effectiveness of the search result clustering method.

Compared with the above "flat" clustering methods, Lawire focused on building the hierarchy of the clusters[7,8]. A statistical language model was used recursively to identify the topic and subtopics terms, which are then organized into a hierarchy. The improvement of Lawire's work includes the Discover algorithm[9], which decreased computation complexity from square to linear. Another aim of Discover was to maximize the coverage while maintaining the distinctiveness of the topics.

In Reference [10], the traditional method the unsupervised problem was transferred to a supervised problem: the salient phrases ranking problem. First, some candidate phrases for queries are labeled by humans, and work as training data to learn some regression models. Then, the models are used to give scores for phrases in the retrieved pages. Finally, by associating documents with the phrases and ranking the phrases, the clusters are formed.

3.2.2 Clustering Presentation and Book Shopping

Searching Shakespeare's "Romeo and Juliet" in an online bookstore would produce over 600 books whose titles contain the search phrase. Clustering-based category presentation can generate seven major categories for displaying the Romeo and Juliet book titles: the original work by Shakespeare (but published by different publishers), abridged version or "Shakespeare-made-simple" versions, notes and teaching guides, commentaries (especially about the tragedy of the story), music titles, theater plays, and "the rest." The original work further contains different publication formats: paperback, hardcover, audio-CD, and computer downloadable versions. Typical categories contain 30 or so items, making it easy to look through any one of them quickly.

In China, a similarly large number of book titles would result from one book search. For example, the search of the Chinese classic "Dream of the Red Chamber" (紅樓夢) in a Chinese online book shopping database produces over 200 books. Four main clusters are formed from the search result: the original work published by different publishers, commentary books, and those that analyze the poems in the novel. (The poems in this novel carry many mysteries and metaphors that become the subject of this category of books.) The fourth cluster, "miscellaneous," includes books on special topics such as the dresses and buildings in the novel.

Mobile phone short messages, however, require an even simpler presentation of such results. Even 30 items in a cluster is too much information to display using short messages. (For mobile phones in China, a short message is limited to 70 Chinese characters.) Therefore, IGP must further select from the list of a cluster those one or two "best" result(s) to display. For book shopping, the choice of the "best" result is based on publisher reputation, availability, price, date of publication, and popularity of the individual books.

3.3 User Input Analysis and Fault Tolerant Search

Incorrect search phrases produce erroneous search results. For example, if the user misspells the word "prairie" with its "sound-alike" version "prayeri," the search for "Little House on the Prayeri" will produce no results or books about churches (house of prayer). If the user was browsing the Web with a broadband-connected PC, he can easily correct his input and search again. But the situation is different in the

mobile environment. Correcting the input and repeating searches can soon become a tedious chore. Researchers of NEC proposed a method helping people reduce error by delivering candidates to the user after some characters are input[11]. Although it is useful on WAP (Wireless Application Protocol), it loses effect in short message. IGP helps prevent such errors by analyzing the user input to determine the most likely search he wants to make. At present, IGP can analyze the Chinese phonetic, Chinese synonym, and ISBN number similarities of the user's input with respect to the information in the category dictionary. (Note that Google's search engine uses a proprietary spelling check technology that can perform certain similar tasks for search inputs in English. In particular, it can correct "Little House on the Prayeri" to "Little House on the Prairie.")

3.3.1 Fault Tolerant Search on Pronunciation

Each Chinese character has an associated "pinyin" representation, which is the phonetic mark of that character. Pinyin is currently the most well-utilized text input method on Chinese PCs and mobile phones. With the pinyin input method, the user types the phonetic marks as alphabets and the PC or mobile phone maps them into Chinese characters. The Chinese mobile phone uses the same alphabets assignment on the numeric keypad as a Western phone.

In the case of pinyin input on a mobile phone, two types of errors often occur. One, due to extraneous or missing key clicks on a mobile keypad, the input Chinese character is different from the one intended. Two, due to the fact that several Chinese characters may have the same pronunciation (or share the same phonetic mark), the resulting Chinese character is different from the one intended. As an example of the latter case, foreign names such as Romeo are normally translated into similar sounding Chinese characters. The generally agreed-upon Chinese translation of Romeo is 羅蜜欧 (Luo-Mi-Ou, or "Satin-Honey-Europe"), but, through the pinyin input method on an ordinary mobile phone, the same key clicks could produce 落密鸥 (Also pronounced Luo-Mi-Ou, but the meaning becomes "Fallen-Secret-Seagull").

IGP resolves this problem by indexing each book title and related category information in pinyin. When a query returns more than one result (book information), pinyin is used to rank the results. The result with the identical pinyin is taken as the correct book title.

3.3.2 Fault Tolerant Search on Synonym

When a user is not very certain about the exact words in a book title, he is likely to input synonyms of those words in the title to begin the search. Because there is no exact match, IGP will find a list of books whose titles are "close" to the user's input. The "closeness" of the match is measured using the synonym (or similar concept) information in IGP's category dictionary.

3.3.3 Fault Tolerant Search on ISBN

Since each book has a unique ISBN (International Standard Book Numbering), it is straightforward to search books by ISBN. But it is easy to make mistakes when copying or inputting the 10-digit ISBN, especially when a numeral repeats like "000." For example, a correct ISBN is "780072959." The user may miss one zero ("0") and input "78072959," or add one extra zero and input "7800072959." When these errors happen, IGP can detect them using the rules of ISBN calculation and the ISBN information in the category dictionary. Then, it returns the book with the correct ISBN.

4. INFRASTRUCTURE OF MOBILE SHOPPING

The infrastructure of mobile shopping consists of four components as shown in **Fig. 2**:

- The Content Provider (CP). The CP provides the essential information of the merchandise, e.g., price. Generally CP is the online bookstore;
- The user who shops on a mobile phone. He

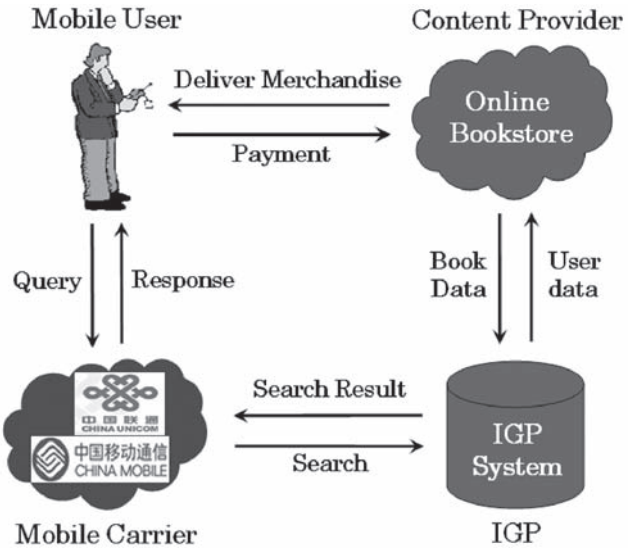

Fig. 2 The infrastructure for Mobile Shopping.

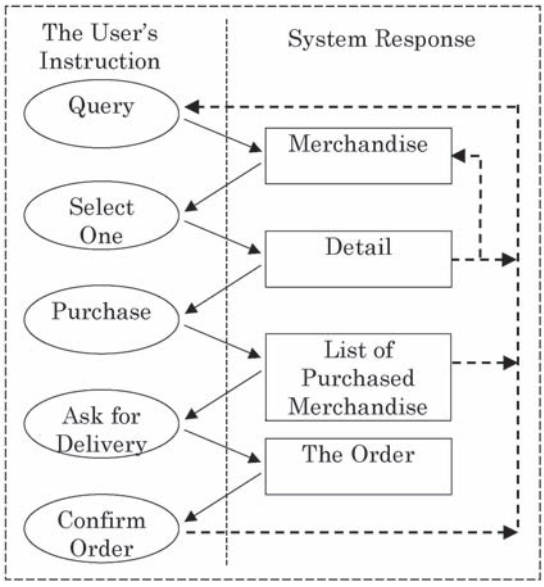

Fig. 3 Short Message Flow for Mobile Shopping.

searches for the merchandise information by short message or WAP. Since short messages are much more popular than WAP in China (there are 320 million short message users versus only 60 million WAP users), short messages are used as an example in the following discussion;

- The Mobile Carrier(MC), for example, China Mobile and China Unicom. The MC transmits the query and system response via its authorized service providers between the user and IGP system;
- The IGP system. It sits between the CP and MC. IGP's interacts with them in two ways. First, IGP continually updates its information category dictionary with the CP's latest book information. Second, each time the user sends a find-book query through the MC to IGP, IGP finds the requested information from the category dictionary and the CP's book database and informs the user. The CP then interacts with the user to complete the purchase.

In China, online purchases are popularly paid COD (Charge On Delivery); the buyer pays cash to the online bookstore's delivery person after he receives and inspects the ordered books.

Figure 3 shows the flow of the mobile shopping process with short messages. **Figure 4** shows an actual shopping process using the mobile-shopping system.

Fig. 4 An actual shopping process.

5. CONCLUSION

In this paper, we introduced the IGP (Information Gathering and Presenting) technology and its application in mobile shopping. IGP represents a new approach to information searches. For a given user query, it aims to gather comprehensively all relevant information and create a precise and concise result for that query. For a subject area, such as books, IGP infers from the Web as many information categories as possible. The categories are then filled with specific values for the subject area's individual instances (e.g., books). The presentation part of IGP utilizes this organized information to enhance the correctness of the user query and create a precise and concise search result for the query.

IGP for mobile book shopping is currently under market test in the Beijing area in China.

REFERENCES

[1] S. Brin and L. Page, "The Anatomy of a Large-scale Hypertextual Web Search Engine," *In Proceedings of the 7 International World Wide Web Conference*, Brisbane, Australia, **3**, ACM Press, 1997.

[2] T. Berners-Lee, J. Hendler and O. Lassila, "The Semantic Web. Scientific American. May issue," 2001.

[3] P. Willett, "Recent Trends in Hierarchical Document Clustering: A Critical Review," *In Information Processing & Management*, **24**, 5, pp.577-597, 1988.

[4] O. Zamir and O. Etzioni, "Web Document Clustering: A Feasibility Demonstration," *In Proceedings of the 19th International ACM SIGIR Conference (SIGIR'98)*, pp.46-54, 1998.

[5] M. A. Hearst and J. O. Pedersen, "Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results," *In Proceedings of the 19th Annual International ACM SIGIR Conference (SIGIR'96)*, Zurich, June, 1996.

[6] O. Zamir and O. Etzioni, "Grouper: A Dynamic Clustering Interface to Web Search Results," *In Proceedings of the 8th International World Wide Web Conference (WWW8)*, Toronto, Canada, 1999.

[7] D. J. Lawrie, W. B. Croft and A. Rosenberg, "Finding Topic Words for Hierarchical Summarization," *In Proceedings of the 24th annual international ACM SIGIR conference (SIGIR'01)*, pp.349-357, 2001.

[8] D. J. Lawrie and W. B. Croft, "Generating Hierarchical Summaries for Web Searches," *In Proceedings of the 26th international ACM SIGIR Conference (SIGIR'03)*, pp.457-458, 2003.

[9] K. Kummamuru, R. Lotlicar and S. Roy, "A Hierarchical Monothetic Document Clustering Algorithm for Summarization and Browsing Search Results," *In Proceedings of the 13th World Wide Web Conference (WWW2004)*, pp.658-665, 2004.

[10] H. J. Zeng, Q. C. He, et al., "Learning to cluster web search results," *In Proceedings of the 27th International ACM SIGIR Conference (SIGIR'04)*, pp.210-217, 2004.

[11] H. Kawai, S. Akamine, et al., "Development and Evaluation of the WithAir Mobile Search Engine," http://citeseer.csail.mit.edu/kawai02development.html, 2002.

*Names of products and companies are trademarks or registered trademarks of each company.

Received March 9, 2005

* * * * * * * * * * * * * * * *

Kai ZHAO received his B.E at the Beijing Institute of Technology in 1993, and received his Dr.E degrees in Pattern Recognition and Intelligent Systems from the Institute of Automation, Chinese Academy of Sciences in 1999. He joined NEC Laboratories China in 2003. He currently focuses on Text/Web Mining, and his general interests include Machine Learning and Text Mining.

Hongwei QI received his B.E and M.E degrees in Mechanical Design and Theory from Hebei University of Technology in 1998 and 2001, and received his Dr.E degree in Pattern Recognition and Intelligent Systems from the Institute of Automation, Chinese Academy of Sciences in 2004. He joined NEC Laboratories China in 2004. He currently focuses on Text/Web Mining, and his general interests include Artificial Intelligence, Machine Learning and Data Mining.

Min-Yu HSUEH studied integrated circuits design and computer-aided design at U.C. Berkeley, where he received his B.S, M.S, and Ph.D degrees, all in Electrical Engineering and Computer Sciences. Other than the two years he worked as a researcher at IBM T.J. Watson Research Center and his current work as the managing director of NEC Labs China in Beijing, China, he spent the majority of his career to date in Silicon Valley where he co-founded and ran new companies, including Cadence Design Systems, currently the world's largest electronic design automation company. His current research interest includes multimedia information retrieval, wireless networking, and high-privacy security systems.

* * * * * * * * * * * * * * * *