A Guided Approach for Personalized Information Search and Visualization

Wei-Bang Chen, Yufeng Li, Seng-Jaw Soong,
and Dongquan Chen
University of Alabama at Birmingham, Birmingham, AL, USA

mybob@ms6.hinet.net yufeng.li@ccc.uab.edu sjsoong@uab.edu
dongquan@uab.edu (corresponding)

Abstract

Our earlier study (Chen, Orthner, & Sell, 2005) showed that it is possible to record the searching strategy, to search and retrieve literature information automatically, and to visualize the retrieved information through a Web interface. The objective of this study was to update the current system into a prototype of the Literature and Information Tracking System to search, retrieve and visualize information in a guided and customized manner. The implementation of the system showed that users could update their account information and searching strategies including topics, online sources, etc., over Web interfaces. The guided approaches for optimized terms usage enhanced the recall, precision, and efficiency. The web-based system may become an efficient tool for both bench scientists and clinicians for their daily information retrieval and visualization. The implemented digital certification kept both convenient access and the highest possible level of security for the retrieved information. The system has great potential to serve a large number of concurrent users.

Keywords: literature search, MeSH, Medline, computer-assisted, information system, system security, Web-based system.

Introduction

Medical Subject Headings (MeSH) and Quality of Information Retrieval

We applied software to use the stored information to search designated information sources such as the medical subject heading (MeSH)-Indexed MEDLINE. The retrieved information was stored on a server and visualized through the Web. The semi-automatic nature of the system still requires the system administrator or a librarian to modify the searching strategy including manually matching the searching terms with MeSH terminology (Chen et al., 2005). Based on our earlier proof-of-principle study, where we showed the possibility that individualized searches could be stored in a database, updated through a system administrator and reused through a software agent, we explored the possibility of offering a guided application of MeSH, thus eliminating the requirement of assistances from the system adminis-

Editor: Kathy Lynch

trator or a librarian. The application of MeSH terms in the right combination always increases recall and sometimes increases precision (Gotzsche & Lange, 1991; Hersh & Hickam, 1992; McKibbon, Friedman, & Friedman, 2002; Wong, Wilczynski, & Haynes, 2004). A recent study also showed that it is possible to enhance the quality of information retrieved by ranking the relevancy of the articles with certain topics such as stem cells (Suomela & Andrade, 2005). Additional training is needed, however, for topics other than stem cells. Our objective is to save searching time for investigators by automating the searching procedure and applying a guided approach to use MeSH terms. Promoting use of MeSH may increase the recall and information retrieval (IR) efficiency, since that the Medline is indexed through the MeSH terms.

Individualized Searching and Retrieving

Individualized information retrieval (IR) is desired by both bench scientists and clinicians to increase the efficiency of their daily information acquisition. We have developed a semi-automatic online search and visualization system in which a system administrator records, parses, and codes the searching strategy in scripts and stores them in a database. The system administrator must, however, manually record all searching terms and parse the terms into an executable program for an automated search. In addition, a user's searching topics or strategy may change over time; thus it makes a manual update difficult for a system administrator, especially when many people are using the system. A web interface that allows users to update their searching terms, schedule, and information sources, etc., is needed for an efficient system.

MyEZbutton: One Place for All Daily Information

In this study, we tried to develop a web interface that allows users to register, update profiles and searching strategy, and view retrieved contents over the web with minimum intervention from a system administrator, as shown in Figure 1. Three major sources of information were included during the prototype development: MeSH-indexed Medline for scientific literature, various websites for personal interests, and a clinical trial database for privileged information such as patient accrual during a certain time frame, race, gender, etc., which needs to be guarded with a higher level of security. A preliminary evaluation was conducted and results analyzed after the prototype development and implementation.

Methods and Implementations

Web Interface Design and User Account Management

As shown in **Figure 1**, the web interface to recruit new users was designed to collect user identifications (IDs), passwords, and registration codes. The registration information will be emailed to the users after registration and will be used to verify the membership. An automatically generated temporary password and/or ID will be emailed to users in case they forget their password and/or ID. The interfaces for strategy information such as topics, MeSH terms, journals, and schedule for searching, etc., were created. Programming languages such as active server pages, HTML, JavaScript and others were used.

Selection of MeSH Terms for Medline, Calculation of Recall and Precision

Each user has four topics for searching MEDLINE and four topics for searching other sources such as a website or a clinical database. Each topic involves up to four MeSH terms connected by Boolean expression (and/or/not). Up to eight journals that were identified with their ISSN numbers were allowed for each topic. The users were guided to use MeSH terms for the MED-

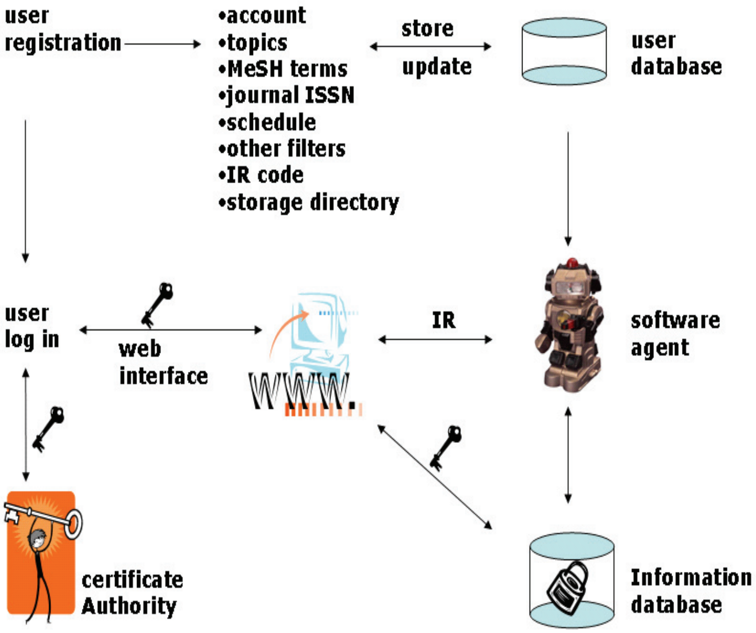


Figure 1. System architecture, data collection, and information flow

The Literature & Information Tracking System starts with user registration. An email address is required to verify the user account and will be used to send the user a password automatically when it is lost. The individualized web interface will be presented with retrieved information after user login as shown in the Figure 4.

LINE search. This is achieved by a suggestion pop-up window after a term was entered. The contents from the PubMed MeSH database were extracted and presented to the user. Different terms should be used in various combinations to generate the best results (high recall and a reasonable precision). Then the optimized MeSH terms and combination were collected in the database: personalized domain specific vocabulary (PDSV) database (Chen et al., 2005). The recall and precision were calculated using the same methods as reported. Briefly, the ***Recall*** is the percentage of total number of retrieved relevant documents within the database (e.g., MEDLINE) by a query. The ***Precision*** is the percentage of relevant documents retrieved among all retrieved document by a specific query or search. Theoretically, it is impossible to get a real number of all relevant documents by any single query. Thus, the combined number from both the original search and the MeSH-assisted search were used to simulate the closest scenario.

Statistical Analyses

A nonparametric method was applied to test the percentage differences in both recall and precision between a regular search and a system-assisted search. P-value was obtained based on a Kruskal-Wallis Test.

Data Sources and Secure Web Server

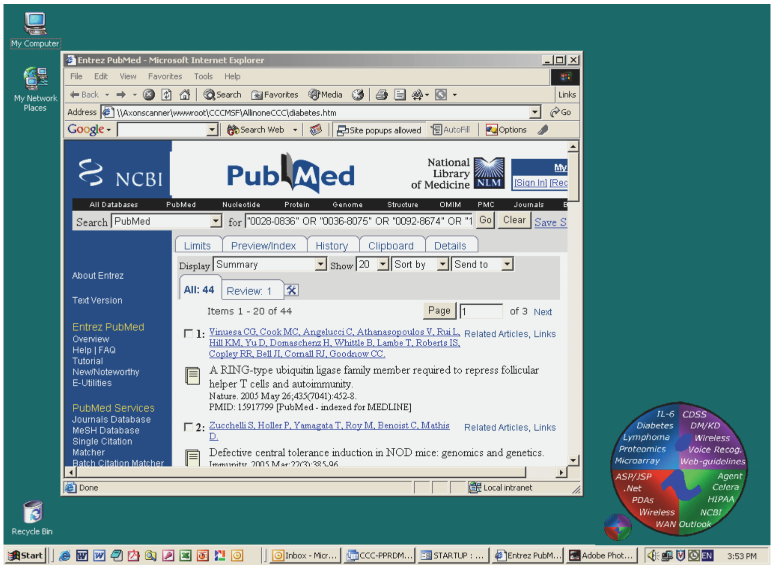
For the information that is retrieved from the clinical trial database, additional security such as user authentication was achieved through user account management and digital certification as reported earlier (Chen, Chen, Soong, Soong, & Orthner, 2004). Briefly, a secure web server that stored all retrieved information was enabled through digital certification. The certificate was issued through a standalone self-managed certificate authority (CA). The user needs a client certificate to access the stored information through a secure socket layer (SSL) over http (https). The self-managed CA issues certificates to both the web server and authorized users. The digital certificates by default were based on secure hash algorithm (SHA)-1 algorithm with key length 512 bits. Added security can be achieved by mapping the certificate with a user account within the Windows-based active directory.

Storage and Agent Software

We adopted an agent software strategy (Baujard, Baujard, Aurel, Boyer, & Appel, 1998; Boyer et al., 1997; Chen, Soong, Grimes, & Orthner, 2004; Gao & Wang, 2004) to search various information sources automatically. The objective was not to develop new agents but to use available ones for searching and retrieving with minimum human intervention. Considering that the retrieved information is usually stored either locally or remotely, we deployed the agent software either on a local workstation (desktop version) or remotely on a server (web version).

Web Version and Desktop Version of the System

The web system retrieves and stores information on the server. The software agent was installed either on a local machine, from where the retrieved files are uploaded onto the server, or it installed on a server, where the retrieved information is also stored. The registered users access the information over the web after logging in. The desktop system has the agent software installed locally as shown in Figure 2. It will search, retrieve, and store information locally in the user's computer. In either case, the agent software automatically executes the searching procedure. The stored information is presented through either desktop buttons or a web browser. The desktop buttons are enabled through a linkage between a desktop image and retrieved information on the local hard drive. The stability and efficiency of the system were tested by repeating the same searching procedures continuously. The time spent was recorded and used for a simulation of a large number of users.



Address \\Axonscanner\wwwroot\CCCMSF\AllinoneCCC\diabetes.htm

Search PubMed for "0028-0836" OR "0036-8075" OR "0092-8674" OR "1

Display Summary Show 20 Sort by Send to

Items 1 - 20 of 44 Page 1 of 3 Next

1: Vinuesa CG, Cook MC, Angelucci C, Athanasopoulos V, Rui L, Hill KM, Yu D, Domaschenz H, Whittle B, Lambe T, Roberts IS, Copley RR, Bell JI, Cornall RJ, Goodnow CC. Related Articles, Links

A RING-type ubiquitin ligase family member required to repress follicular helper T cells and autoimmunity.
Nature. 2005 May 26;435(7041):452-8.
PMID: 15917799 [PubMed - indexed for MEDLINE]

2: Zucchelli S, Holler P, Yamagata T, Roy M, Benoist C, Mathis D. Related Articles, Links

Defective central tolerance induction in NOD mice: genomics and genetics.
Immunity. 2005 Mar;22(3):385-96.

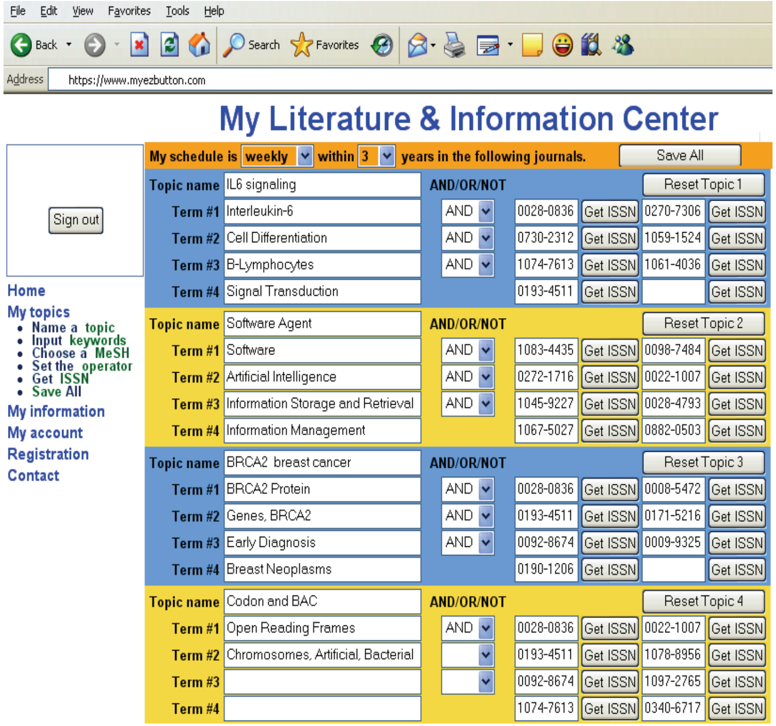
Figure 2. The desktop version of the system.

The system has an image map that links each of the topics labeled on the image with linkages to the retrieved information stored in the local hard drive. Clicking of any part of the image opens up an Internet Explorer browser to display the linked information for the related topics. The image can be customized or individualized.

Results and Discussion

Online Searching Strategy

The terms/words selection is critical for both recall and precision in a search. Currently, the search engine in PubMed automatically matches the entered terms with MeSH and will use both the matched and unmatched terms for a search. The unmatched terms will be used against the title and abstracts. In our system, as showed in Figure 3, after the MeSH terms, schedule, journal international serial number (ISSN) are entered and saved, the software agent will search the Medline according the schedule and the retrieved linkages will lead the users to the visualization interface as showed in Figure 4.



Address https://www.myezbutton.com

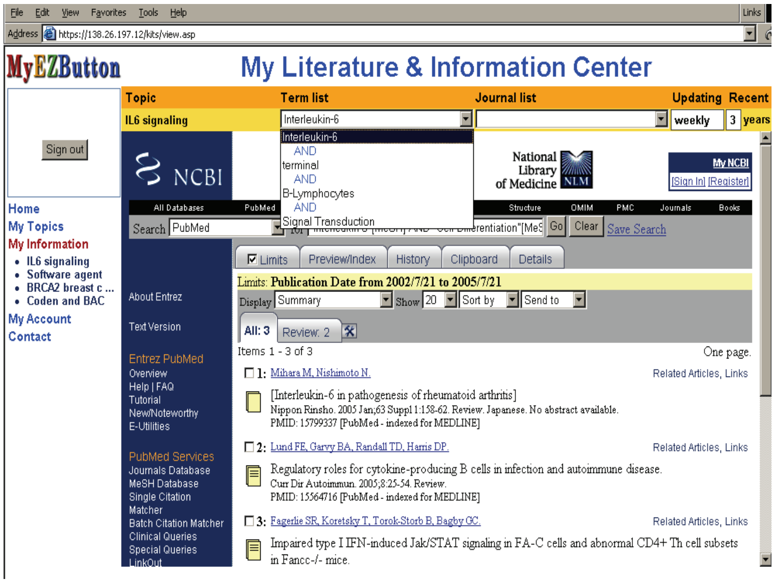
My Literature & Information Center

My schedule is weekly within 3 years in the following journals.

		AND/OR/NOT		
Topic name	IL6 signaling			
Term #1	Interleukin-6	AND	0028-0836	0270-7306
Term #2	Cell Differentiation	AND	0730-2312	1059-1524
Term #3	B-Lymphocytes	AND	1074-7613	1061-4036
Term #4	Signal Transduction		0193-4511	
Topic name	Software Agent			
Term #1	Software	AND	1083-4435	0098-7484
Term #2	Artificial Intelligence	AND	0272-1716	0022-1007
Term #3	Information Storage and Retrieval	AND	1045-9227	0028-4793
Term #4	Information Management		1067-5027	0882-0503
Topic name	BRCA2 breast cancer			
Term #1	BRCA2 Protein	AND	0028-0836	0008-5472
Term #2	Genes, BRCA2	AND	0193-4511	0171-5216
Term #3	Early Diagnosis	AND	0092-8674	0009-9325
Term #4	Breast Neoplasms		0190-1206	
Topic name	Codon and BAC			
Term #1	Open Reading Frames	AND	0028-0836	0022-1007
Term #2	Chromosomes, Artificial, Bacterial		0193-4511	1078-8956
Term #3			0092-8674	1097-2765
Term #4			1074-7613	0340-6717

Figure 3. The Web version of the Literature and Information Tracking System

The web version interface enables users to register and records the searching topic and strategies using Boolean expression to focus the search. MeSH terms will be suggested after a term is entered. ISSN numbers are also suggested after a journal name is entered. The "Save All" button sends all collected information to a database (Chen et al., 2005). The "My information" button leads to stored information that was updated regularly according to the user's schedule. The user interface was protected by login information and a Windows-based user account. "My account" allows users to update the topics, terms, journals, schedules, and password, etc.



My Literature & Information Center

Topic	Term list	Journal list	Updating	Recent
IL6 signaling	Interleukin-6		weekly	3 years

Interleukin-6
AND
terminal
AND
B-Lymphocytes
AND
Signal Transduction

Search PubMed for ...erentiation"[MeS

Limits: **Publication Date from 2002/7/21 to 2005/7/21**

Display Summary Show 20

All: 3 Review: 2

Items 1 - 3 of 3 One page.

□ 1: Mihara M, Nishimoto N.
[Interleukin-6 in pathogenesis of rheumatoid arthritis]
Nippon Rinsho. 2005 Jan;63 Suppl 1:158-62. Review. Japanese. No abstract available.
PMID: 15799337 [PubMed - indexed for MEDLINE]

□ 2: Lund FE, Garvy BA, Randall TD, Harris DP.
Regulatory roles for cytokine-producing B cells in infection and autoimmune disease.
Curr Dir Autoimmun. 2005;8:25-54. Review.
PMID: 15564716 [PubMed - indexed for MEDLINE]

□ 3: Fagerlie SR, Koretsky T, Torok-Storb B, Bagby GC.
Impaired type I IFN-induced Jak/STAT signaling in FA-C cells and abnormal CD4+ Th cell subsets in Fancc-/- mice.

Figure 4. My Information: The individualized Web interface

The "My information" button leads to preview windows of all topics, searching strategy such as MeSH terms, journal names, schedule, etc. The preview panel shows a list of literature retrieved from MEDLINE. Clicking on the link leads to the article abstract in MEDLINE.

We compared both recall and precision of four different topics between the direct searches and system-guided and MeSH-based searches. Among these typical searches, the authors are considered experts in these fields and are thus reliable in the judgment of relevancy. In all topics, the original searches generated fewer recalls and usually higher precision as shown in Tables 1-4. The system-assisted and MeSH-based searches generated much higher recall: ↑97% (topic 1), ↑95 % (topic 2), ↑52% (topic 3), and ↑1-5% (topic 4). In all four topics, searches with original terms generated a relative higher precision. A higher precision is not desired, however, if only a small percentage of relevant documents are retrieved (3% in topic 1 and 2% in topic 2, e.g.). To estimate the recall-precision relations, we combine both the regular search and the system-assisted searches and graphed in Figure 5. We observed the trend of a tradeoff between the recall and precision, in general. As mentioned earlier, although the real number of relevant documents may be much higher than the combination of relevant documents in the two searches, the strategy of combining the two is a close estimate for relevant documents.

Table 1. Topic 1: IL-6-induced B cell terminal differentiation

	Original terms	MeSH terms	↑ or ↓
Term 1 and	IL-6-induced	Interleukin 6	
Term 2 and	B cell	B-lymphocytes	
Term 3	terminal differentiation	cell differentiation	
recall	5/176 (3%)	176/176 (100%)	↑97%
precision	5/5 (100%)	176/296 (59%)	↓41%

Table 2. Topic 2: Software agent-based IR and knowledge management

	original terms	MeSH terms	↑ or ↓
term 1 and	software agent-based	software artificial intelligence	
term 2 and	information retrieval	information storage and retrieval	
term 3	knowledge management	Information management	
recall	2/79 (2%)	77/79 (97%)	↑95%
precision	2/2 (100%)	77/122 (63%)	↓37%

Table 3. Topic 3: BRCA2 in early detection of breast cancer

	original terms	MeSH terms	↑ or ↓
term 1 and	BRCA2 in	BRCA2 protein OR genes, BRCA2	
term 2 and	early detection of	early diagnosis of	
term 3	breast cancers	breast neoplasms	
recall	27/76 (36%)	67/76 (88%)	↑52%
precision	27/28 (96%)	67/72 (93%)	↓3%

Table 4. Topic 4: Search for coding sequences among bacterial artificial chromosome (BAC) clones

	original terms	original terms	MeSH terms	↑ or ↓
term 1 and	coding sequences	ORFs	open reading frame	
term 2	BAC clones	BACs	chromosomes, artificial, bacterial	
recall	32/80 (40%)	35/80 (44%)	36/80 (45%)	↑ 1-5%
precision	32/32 (100%)	35/37 (95%)	36/42 (86%)	↓ 9-14%

Note: some bacteria naming convention such as *Bac. Stearothermophilus*, etc., and some words such as tobacco or baculovirus may interfere the search when “bac” is used.

It is also observed that some words with the same meaning but different format (abbreviation, e.g.), may return vastly different documents (topic 4, e.g.). The underlying reason may be that the index in MEDLINE does not cover all possible biological meanings of a term such as BAC/BAC clones/chromosomes, artificial, bacterial; or those bacteria naming conventions such as Bac. Stearothermophilus, interferes with the search for the word BAC (bacterial artificial chromosome).

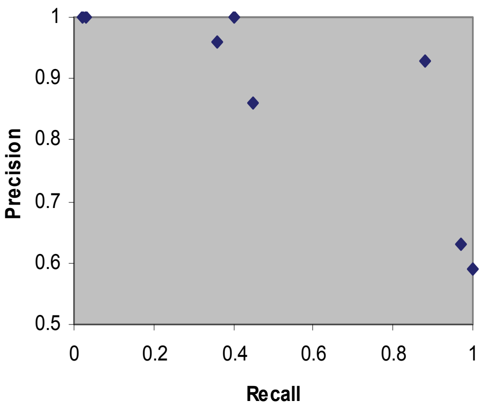


Figure 5. The recall-precision relations

To illustrate the recall-precision relations, all searches including both the original and the system-assisted searches are included.

To further compare recall and precision between the original and the system-assisted searches, we combined data from Tables 1-4. A statistical comparison using a nonparametric method was performed. As shown in Table 5, the system-assisted search has a median percentage recall of 92.5% versus 19% in the original search (p=0.0209). Although the original search has a higher median percentage precision of 98% versus 78% in the assisted search (p=0.0180), it did not return the majority of the relevant documents.

Table 5. A comparison of a regular vs. a system-assisted search

	Recall*		Precision**	
	regular search	assisted search	regular search	assisted search
Topic I	5/176 (3%)	176/176 (100%)	5/5 (100%)	176/296 (59%)
Topic II	2/79 (2%)	77/79 (97%)	2/2 (100%)	77/122 (63%)
Topic III	27/76 (36%)	67/76 (88%)	27/28 (96%)	67/72 (93%)
Topic IV	32/80 (40%)	36/80 (45%)	32/32 (100%)	36/42 (86%)

- p=0.0209 based on a Kruskal-Wallis test. ** p=0.0180 based on a Kruskal-Wallis test.

Based on these observations, we concluded that the different searching strategies may significantly change the search outcomes and we suggest the use of MeSH terms to increase recall with a reasonable precision, if not higher, to increase the IR efficiency and overall performance.

Agent Software for Automatic Searching, Retrieving, and Storage

To test the capacity, stability, and efficiency of the system, we created searching tasks that involved different numbers of users. The software was programmed to finish all IR procedures / tasks in a fixed length of time, which is adjustable and optimized to complete the search procedure for all users successfully. By repeating the same searching procedures continuously, the system was tested for its stability to complete the searching procedures successfully.

The agent software took 15 minutes to finish a search procedure that was designed for 40 users (4 topics each and 160 searches total from various websites including the MEDLINE). A searching task that involved a larger number of users (4,000, e.g.) was simulated by concatenating the procedure for 40 users 100 times. As shown in Table 6, less than 5% of tasks run uncompleted. The technical problems include among others that the server is unavailable; computer freezes; system memory is full; temporary internet folder is full. Continued optimization of the program and ever-increasing hardware and network capability may further decrease the uncompleted percentage of search procedures and increase the number of users the system could serve.

Table 6. System efficiency and stability

	time (hrs)*	completed (%)
40 users	0.25	100
400 users	2.5	98
4,000 users	25.0	95

Note: the programmed time to finish the searches.

The system may run with higher efficiency. Our test was based on a laptop computer running as a software agent. A dedicated server with faster central processing unit (CPU) and larger memory may further improve the system performance and minimize the problems mentioned above. In addition, considering that not all users may have four topics, that the continuously increased computer performance, and that not all users are online at the same time, it is safe to say that the system may handle 3,000-4,000 users for daily updated searches.

Access Control and Security Enhancement over the Clinical Information

Access control is necessary only for the stored clinical information or any user information considered private. The digital certificates, by default, were based on secure hash algorithm (SHA)-1 with a key length of 512 bits. More vigorously encrypted algorithms such as the SHA-256/512, the message digest (MD), the message authentication code (MAC), a combination of various algorithms such as the keyed-hash message authentication code (HMAC) and the digital certificate encrypted with a longer key should be tested and implemented over time to maintain the highest possible security.

To maintain security and at the same time to provide convenience, the system could be viewed through wireless connection (data not shown) as reported before (D. Chen et al., 2004). The system automatically sends users ID and password upon registration and resends when forgotten or requested. To monitor the system access, a system log is enabled and audited regularly. The system requirement, access, and availability are listed in the Appendix.

The Web versus the Desktop Version of the System

The web version system runs on a server and the desktop version on a local workstation or a laptop computer. In the web version system, the agent software runs either on the server or locally, but uploads the searching results onto the server. In the desktop system, as shown in **Figure 2**, the agent software runs locally and store results locally in a hard drive. The user accesses the retrieved information through an active desktop map as shown in **Figure 1.** We found that the desktop system performs similarly since the performance depends on searching strategy, computer and network speed and not the locality.

Conclusion

The web-based Literature and Information Tracking System is an efficient tool for both bench scientists and clinicians for their daily information needs. Our system adds more functions and is more user-friendly than our previously reported system. Both the web and the desktop approaches help increase recall and overall performance of information retrieval through a guided application of MeSH. Digital certification keeps both convenient access and the highest possible security for the retrieved information especially those from privileged information systems such as a clinical trial information system. The computer-guided approach and the self-updatable nature of the searching strategy including terms, Boolean expressions, journals etc., minimize the efforts and intervention for system administration, thus having a great potential to serve a large number of users at the same time.

References

Baujard, O., Baujard, V., Aurel, S., Boyer, C., & Appel, R. D. (1998). Trends in medical information retrieval on Internet. *Computers in Biology and Medicine, 28*(5), 589-601.

Boyer, C., Baujard, O., Baujard, V., Aurel, S., Selby, M., & Appel, R. D. (1997). Health On the Net automated database of health and medical information. *International Journal of Medical Informatics, 47*(1-2), 27-29.

Chen, D., Chen, W.-B., Soong, M., Soong, S.-J., & Orthner, H. F. (2004). *A Web-based system for clinical trials: to turn Access into a web-enabled secure information system*. Manuscript submitted for publication.

Chen, D., Orthner, H. F., & Sell, S. M. (2005). Personalized online information search and visualization. *BMC Medical Informatics and Decision Making, 5*(1), 6.

Chen, D., Soong, S. J., Grimes, G. J., & Orthner, H. F. (2004). Wireless local area network in a prehospital environment. *BMC Medical Informatics and Decision Making, 4*(1), 12.

Gao, G. Y., & Wang, S. K. (2004). A multi-agent system architecture for geographic information gathering. *Journal of Zhejiang University Sciences, 5*(11), 1367-1373.

Gotzsche, P. C., & Lange, B. (1991). Comparison of search strategies for recalling double-blind trials from MEDLINE. *Danish medical bulletin, 38*(6), 476-478.

Hersh, W. R., & Hickam, D. H. (1992). A comparison of retrieval effectiveness for three methods of indexing medical literature. *The American journal of the medical sciences, 303*(5), 292-300.

McKibbon, K. A., Friedman, P. W., & Friedman, C. P. (2002). Use of a MeSH-based index of faculty research interests to identify faculty publications: An IAIMSian study of precision, recall, and data reusability. *Proceedings of American Medical Informatics Association Symposium*, 514-518.

Suomela, B. P., & Andrade, M. A. (2005). Ranking the whole MEDLINE database according to a large training set using text indexing. *BMC Bioinformatics, 6*, 75.

Wong, S. S., Wilczynski, N. L., & Haynes, R. B. (2004). Developing optimal search strategies for detecting clinically relevant qualitative studies in MEDLINE. *Medinfo, 11*(Pt 1), 311-316.

Appendix

Availability and Requirements

The system is currently available at http://138.26.197.12/kits/default.asp. The user manual is provided on the site. With minimum effort, the system can be implemented with the capacity in the Table 6. We welcome user registration, feedback, and comments on the system, which requires

Widows operation system and no specific type of browser. This will further improve the system for its convenience, security, and performance.

Biographies

Dr. **Dongquan Chen** (Biostatistics and Bioinformatics Unit, Comprehensive Cancer Center and Department of Health Services Administration, School of Health Related Professions, University of Alabama at Birmingham) is the principal investigator of the project. He was trained as a molecular biologist with Bioinformatics postdoctoral experience. He is an Assistant Professor and an Associated Scientist in the Comprehensive Canter Center of the University of Alabama at Birmingham, USA. He was award an Individual Postdoctoral Fellowship from the US National Library of Medicine. His training in Medical Informatics earned him a Master's degree in Health Informatics. He has developed several secure information systems to support research scientists in both Bioinformatics and Medical Informatics fields.

Mr. **Wei-bang Chen,** a PhD candidate, (Department of Computer and Information Sciences. School of Natural Sciences and Mathematics, University of Alabama at Birmingham), contributed most of the programming for web interfaces design, database administration, and implementation of the system security.

Dr. **Yufeng Li** (Biostatistics and Bioinformatics Unit, Comprehensive Cancer Center, University of Alabama at Birmingham) did the statistical analysis for the study

Dr. **Seng-Jaw Soong** (Biostatistics and Bioinformatics Unit, Comprehensive Cancer Center, University of Alabama at Birmingham) is an advisor to the project and provided the financial support.

All contributed to the project planning, coordination, and performance evaluations.