SEACOIN – An Investigative Tool for Biomedical Informatics Researchers

Eva K. Lee, PhD[*,1,2,3,4], Hee-Rin Lee, MS[1,2,5], Alexander Quarshie, MD, MS [6,7,8]

[1]Center for Operations Research in Medicine and HealthCare, [2]NSF I/UCRC Center for Health Organization Transformation, [3]Cemter for Bioinformatics and Computational Genomics, [4]School of Industrial and Systems Engineering, [5]Digital Media, School of Literature, Communication, and Culture, Georgia Institute of Technology, Atlanta, GA; [6]Biomedical Informatics Unit, [7]Biostatistics and Data Management Core, [8]Clinical Research Center, Morehouse School of Medicine, Atlanta, GA

Abstract

Peer-reviewed scientific literature is a prime source for accessing knowledge in the biomedical field. Its rapid growth and diverse domain coverage require systematic efforts in developing interactive tools for efficiently searching and summarizing current advances for acquiring knowledge and referencing, and for furthering scientific discovery. Although information retrieval systems exist, the conventional tools and systems remain difficult for biomedical investigators to use. There remain gaps even in the state-of-the-art systems as little attention has been devoted to understanding the needs of biomedical researchers.

Our work attempts to bridge the gap between the needs of biomedical users and systems design efforts. We first study the needs of users and then design a simple visual analytic application tool, SEACOIN. A key motivation stems from biomedical researchers' request for a "simple interface" that is suitable for novice users in information technology. The system minimizes information overload, and allows users to search easily even in time-constrained situations. Users can manipulate the depth of information according to the purpose of usage. SEACOIN enables interactive exploration and filtering of search results via "metamorphose topological visualization" and "tag cloud," visualization tools that are commonly used in social network sites. We illustrate SEACOIN's usage through applications on PubMed publications on heart disease, cancer, Alzheimer's disease, diabetes, and asthma.

*****Corresponding author: eva.lee@gatech.edu

Introduction

Peer-reviewed scientific literature is a prime resource for accessing worldwide scientific knowledge, but its continuing growth and diversification demand systematic and automated efforts to effectively utilize the information that it contains[1]. Most researchers use electronic retrieval systems (e.g., PubMed) for their studies, searching for key results in bio/clinical advances, finding what has been done, where the research was conducted, when it was performed, and by whom.

However, current systems present difficulties for biomedical researchers. Since many systems are developed by information experts, they may be too complex for a novice user to properly utilize search query features, and in turn query results may be so voluminous as to overload the user. For example, researchers reported that advanced searches utilizing Medical Subject Headings (MeSH) terms are rarely used although they could retrieve concise search results[2]. Bridging the gap between biomedical researchers and computational linguists is crucial to the success of biomedical literature mining[1]. In previous work, biomedical researchers have offered recommendations in user interface design that will make the tools accessible to non-specialists in informatics[1]. For example, some studies investigated existing biomedical information retrieval systems and reported users' needs by interviewing the users on their needs and analyzing their logs. However, few systems researchers have explicitly defined the needs before they design the interface.

In this study, we first define user needs by searching and organizing previous studies on needs of biomedical specialists. Next, we evaluate previous studies on user electronic biomedical literature retrieval systems by applying the defined needs. Finally, we derive and implement new informatics and visualization design interface based on users' desires. Preliminary validation of usage of our system is performed on data/search analysis using MEDLINE.

Users' Behavior Pattern and Needs

Previous studies have examined biomedical researchers' behavior patterns and their general needs from PubMed, the most popular/common biomedical search engine for the MEDLINE database[2, 3]. Some analyzed query logs to gain insights, while others documented difficulties or limitations of current systems. Davies et al[4] interviewed biomedical experts to better understand their needs. Based on the literature, we identify seven characteristics that represent users' needs on biomedical information retrieval system that focuses on PubMed databases.

1. *Most users prefer short and simple queries based on keywords.* Studies showed that most queries are composed of keywords (94% of searches) [4]. Queries are usually short with very few terms[3]. The average number of tokens per query is 3.54[3]. High-level concept words such as 'cancer', 'cell', 'review', and 'protein' are commonly-used terms instead of specific gene names or formal clinical words as in Table 1[2]. 'Kinase' is relatively the most formal term among 40 top common words.
2. *Users mostly issue queries that need to explore uncharted database.* Only one quarter of queries were navigational (bibliographic query) which try to find a specific paper by employing original search categories (e.g., author, journal

name) of database structure[2]. The remaining queries were informational searches intended to satisfy information needs on the keyword as a topic. An informational query needs to extract data from unstructured format by analyzing and parsing the retrieved literature. For example, when a use enters 'asthma' as a query, they expect the system to retrieve literature whose topic pertains to asthma. Unfortunately, the database searched by the system may not have structured data indicating if each paper is 'asthma related' or not. Most data files are unstructured documents. The system should provide users with the ability to search and analyze abstracts or other unstructured information/documents so as to locate relevant literature.

3. *Users want to retrieve information focused on one category and each user has very diverse interest in different domains.* Dogan at el. reported 60% of queries were annotated with only one category[3]. In other words, users need in-depth search focusing on a category. They would like to acquire depth of literature pertaining to the key topic of their choice rather than different context of literature belonging to similar categories. On the other hand, users issue queries on a large variety of topics without dominant search terms or topics[2].

Table 1. Common terms in informational query[2]

Term	Frequency
cancer	46,370
cell	39,687
review	35,272
disease	21,337
protein	20,417
cells	17,574
human	17,512
receptor	15,837
treatment	15,715
syndrome	15,476

4. *Users want a familiar and simple interface that does not require special knowledge of information technology.* Ironically, although MeSH, which provides tags of possible topics for each literature, allows topic-based search and in-depth search, it is seldom used[2]. One reason is the complex interface of MeSH term search. Users need to know query knowledge in order to use MeSH terms and need to get through several steps to build a query. Clinical and biomedical specialists have voiced that they have difficulties using advanced features of PubMed because of lack of expert knowledge in information technology as well as a lack of time to acquire the knowledge. An intuitive search interface that is tailored to biomedical specialists' search knowledge is important[1].
5. *Users are sensitive and overwhelmed by 'Information overload.'* Voluminous search results have a negative effect on user experience, as the information returned is not conducive for results summarization. In one study, it was stated that the result set contains an average of 13,798 citations with a median of 17 citations[3]. As a result, users are more inclined to issue a new query than to request an abstract or full-text view[3]. Similarly, users are less likely to select a citation as the result returns get bigger[3].
6. *Few users review results beyond the first page.* Over 80% of the clicks for detail views occurred on one of the top 20 citations returned in the result set[3]. That is, most clicks happened on citations in the first result page (by default, PubMed returns 20 results per page). The number of clicks for the documents in the later pages degrades exponentially[3].
7. *Page format has effect on user behavior.* A study that observed the ordinal position of the clicked item found that users are more likely to click the first and the last returned citations of each result page[3]. This suggests that users are influenced by the result page format when selecting returned citations rather than simply following the retrieval order of PubMed.

Based on this analysis, as well as our discussion with clinical and biomedical investigators, we conclude that users desire:

- an input interface that is suitable for keyword-based search;
- retrieve results (based on the keywords they issue as a topic) as a response to informational query;
- considerable depth of returned information focusing on one category from a diverse domain;
- simple-to-use interface that requires no special knowledge in information technology
- return of results without information overload
- display results onto a single page
- page format that takes into account users' behavior and interaction with the system, with focus on top-down and bottom-up display.

Existing Systems: Features, limitations, and Challenges

Below, we summarize how these seven factors related to user needs are reflected in existing tools and systems.

Most existing systems provide a suitable input interface for short and simple keywords. PubNet requires users to create a query by specifying node category (e.g., author, paper, GenBank) and edge category (e.g., co-authorship, MeSH terms) [5]. However, most systems import the PubMed query interface into their applications, which provides a fine-grained input interface for keyword-based query[5-11].

The majority of the systems focus on information retrieval of uncharted territories. Uniformly, most encounter similar difficulties in exploring a full-text database for informational query, and developers put much effort on automatic text analysis to identify and extract relevant facts[9]. PubMed abstracts are used as input and data are explored by document clustering, network representations of the underlying conceptual relationships and the mapping of search results to MeSH and Gene Ontology trees[10].

Although the developers pointed out the importance of in-depth search, and some employed a pre-defined ontology such as MeSH and Gene Ontology trees to organize result sets[7, 9, 10], little attention has been paid to the depth of retrieved information focusing on one category in a diverse domain. Moreover, fixed ontology structures focus on describing how the topic (issued keywords) is related to the predefined categories rather than explaining the topic at different levels. Additionally, a fixed ontology structure, which is generated manually by a human, cannot cover all the domains of biomedical literature. We should note that since 2008, MEDLINE has expanded at a rate of approximately 2,200 new entries per day (calculated from PubMed 2008 indexed entries) [12].

Many systems tend to have a much more complex interface than users want. Although developers strive to create intuitive and easy-to-use interfaces, few consider users' experience and knowledge on information retrieval technology. For example, developers utilize network visualization to represent the associations of biomedical entities in the retrieved literature without considering its complexity with overlapped nodes[5, 9].

Very often, the systems provide an overwhelming amount of filtering functions and choices. Some enable filtering of results by selecting a category in a fixed ontology [9, 10] or by selecting important words from the result sets[10]. For example, PubOnto provides four fixed ontology structures, each offering more than 500 categories as optional selections to filter the results[10]. Overlapped nodes of visualization can also contribute to information overload.

The number of result pages is not reflected as an essential design rationale. Systems with visualization such as AliBaba and PubOnto offer a one-page summarization of results[9, 10]. However, most other systems return excessive number of pages as in PubMed's. When the word 'cancer' is searched, most systems return more than 100,000 pages of information.

Few systems examine the format of a result page and its impact to user's response. Little effort has been devoted to arranging components in a page. A number of systems follow the format of PubMed such as Anno O'Tate although it has been commented that its format needs improvement. Systems with visualization present suitable page format by arranging a component containing results on the bottom and thus making it easy for users to access the final information.

Table 2 Comparison between text-based applications and applications with visualization

	Text-based system/tool	System/tool with Visualization
Depth	Dissatisfied	Dissatisfied
Familiar/simple-to-use interface	Satisfied	Dissatisfied
Information overload	Dissatisfied	Dissatisfied
# of returned pages	Dissatisfied	Satisfied
Page format	Dissatisfied	Satisfied

Contrasting existing systems and tools against the desires of users, we identify dissatisfaction gaps in five areas: depth of retrieved information, familiar and simple-to-use interface, information overload, number of returned pages, and design of result page format that is conducive to users' understanding of results. The gaps are different depending on the format of the tools — whether it is text-based or with visualization features (Table 2). Systems with visualization tend to satisfy the needs on 'number of result pages' and 'page format' by summarizing an enormous number of citations into a graph, where some effort has been expended on content arrangement. On the other hand, text-based systems satisfy the needs on familiar and easy-to-use interface by employing a general Internet interface such as hypertext link. However, all systems lack in the depth of retrieved information. Although some systems with visualization employ a predefined ontology, they do not offer different depths of information. Information overload is seen either with enormous amount of text results returned, or that the information is entangled into an enormous visual data format. Most systems try to provide as many functions as possible without articulating what functions are actually needed and for what purpose.

SEACOIN

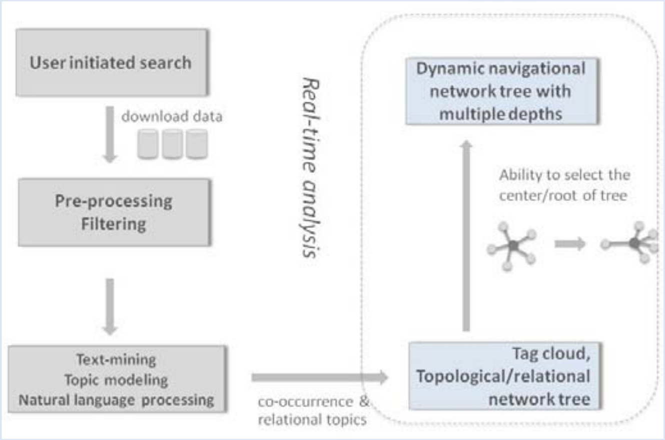
Search **E**xplore **A**nalyze **CO**nnect **IN**spire, SEACOIN, is a web-based utility that attempts to alleviate the shortcomings of existing search technology. SEACOIN's design is based on needs of biomedical researchers. The design addresses the seven user-defined needs previously described as well as four problems in visualization methods. It utilizes word cloud and metamorphoses topological visualization concepts. Briefly, SEACOIN allows

i. keyword-based queries through the input interface;
ii. uncharted territories analysis (results are parsed by machine learning and indexed searching techniques and are subsequently summarized into word cloud and topological visualization constructs);
iii. topological visualization that contains multiple levels of depth;
iv. use of tag clouds, one of the most familiar visualization tools that can increase user accessibility significantly;
v. exploration of uncharted database;
vi. dynamic 1-page summarization and incorporation of users' behavior into effective page format and content visualization.

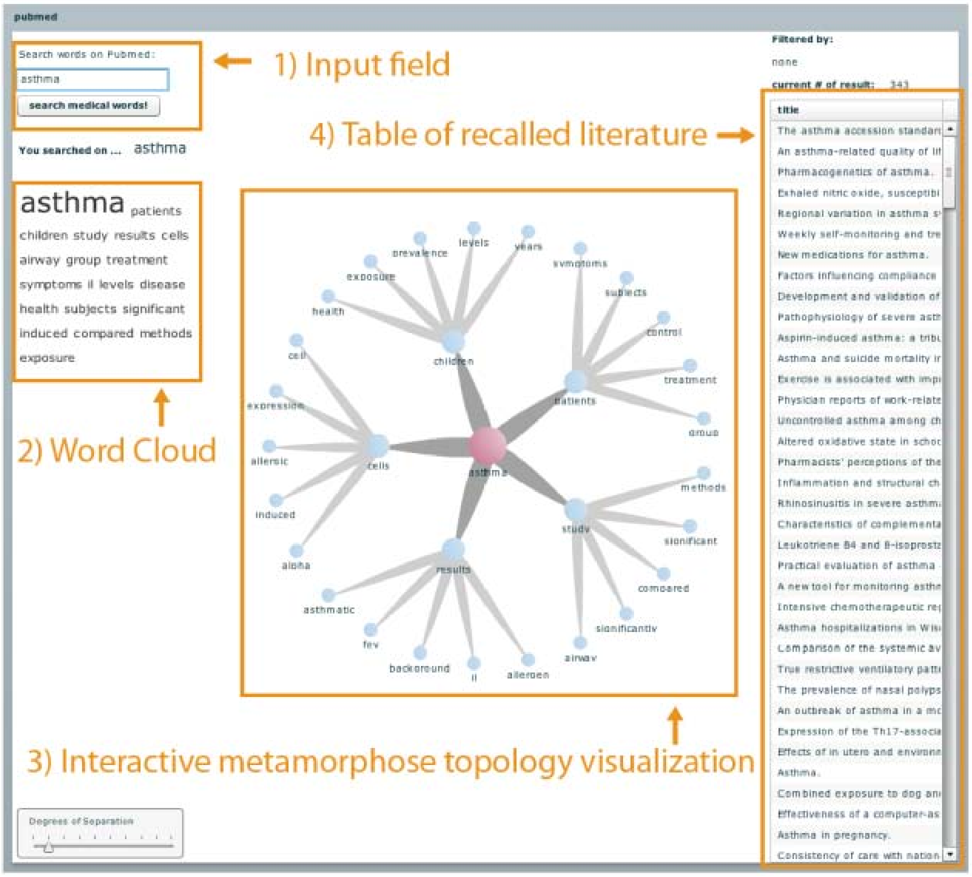
Figure 1 shows the schema workflow within the SEACOIN investigative engine.

SEACOIN is designed for use by biomedical researchers whose time is limited and who can benefit from a simple-to-use visualization-based search tool that can present the results of their queries in an easy-to-understand and content-driven summarized format. When a user initiates a search through key words, SEACOIN begins the real-time search process, e.g., in our examples below, we search the PubMed databases. The related data are pulled, pre-processed, and filtered. Text mining, topic modeling, and natural language processing are performed on these unstructured texts. Co-occurrence and relational topics are identified that will then be utilized to establish the tag clouds and topological/relational network tree for eventual dynamic navigation. The tree consists of multiple depths, allowing a top-down and bottom-up approaches to review the data in visualization and user-driven interactive format. The key challenges in the development of SEACOIN are the computational time in analyzing text-based data, the real-time generation of the content relationship, and the resulting navigational capability.

Figure 1. A simplified schematic workflow within SEACOIN investigative tool



The SEACOIN summarization screen includes 1) input field, 2) word cloud, 3) interactive metamorphose topology visualization with multiple-depth navigation, and 4) table of recalled literature, displayed in four regions as shown in Figure 2. Specifically, the upper left input field accepts queries to PubMed. Under the input field, word cloud shows a quick overview of results from parsing the abstracts returned for the query. The interactive graph in the center contains detailed information of results from mining and parsing the abstracts in real-time. The table on the right includes the list of retrieved literature.



Filtered by:

none

current # of result: 343

title
The asthma accession standar
An asthma-related quality of lif
Pharmacogenetics of asthma.
Exhaled nitric oxide, susceptibi
Regional variation in asthma s
Weekly self-monitoring and tre
New medications for asthma.
Factors influencing compliance
Development and validation of
Pathophysiology of severe asth
Aspirin-induced asthma: a tribu
Asthma and suicide mortality ir
Exercise is associated with impi
Physician reports of work-relate
Uncontrolled asthma among ch
Altered oxidative state in schoc
Pharmacists' perceptions of the
Inflammation and structural ch
Rhinosinusitis in severe asthm:
Characteristics of complementa
Leukotriene B4 and 8-isoprostz
Practical evaluation of asthma
A new tool for monitoring asthr
Intensive chemotherapeutic re|
Asthma hospitalizations in Wis
Comparison of the systemic av
True restrictive ventilatory patt
The prevalence of nasal polyps
An outbreak of asthma in a mc
Expression of the Th17-associa
Effects of in utero and environ
Asthma.
Combined exposure to dog an
Effectiveness of a computer-as
Asthma in pregnancy.
Consistency of care with nation

Figure 2. SEACOIN summarization display panel: 1) input field, 2) word cloud, 3) interactive metamorphose topology visualization with multiple depth navigation, and 4) table of recalled literature.

Input field The input query interface of the current version of PubMed is a popular search tool and is easy to use for biomedical investigators. We therefore import the PubMed query interface into the SEACOIN tool. When a user types in a query, it is sent to PubMed using the NCBI E-Utilities (ESearch and EFetch)[10] to obtain the PubMed IDs. This takes advantage of existing pre-processing that occurs within PubMed[10]. Given the set of PubMed IDs, articles are first searched for in a local database. For articles not found in the local database, E- Utilities are used to download the relevant records directly from PubMed[10]. The local database is established based on recent queries by users, and is updated dynamically as various searches are performed within SEACOIN and information is accumulated.

Word cloud Word cloud is a familiar simple-to-use visualization construct on social network sites. It provides an impression of the person and his or her interests[14]. SEACOIN employs word cloud to provide an impression of all searched literature in a conventional way without information overload. The size and order of each tag represent frequency in recalled literature. Clicking a tag alters the center of topology visualization and re-arranges the structure of topology visualization centered on the clicked item. Also, the table of recalled data is updated and filtered based on the clicked word of word cloud.

Users can change the center/root of the graph and the degree of separation. It allows the navigation of content at the user's discretion. Figure 3 illustrates the topological network of contents. The upper left figure is the initial default tree with the chosen word as the center of the tree. The upper right figure displays the entire tree with the available level of depths (4 levels are shown here). The lower left figure shows a rearranged graph by changing its center node to a node of level 2 with degree value 2 (simply by clicking on that node). The lower right figure shows a rearranged graph by changing its center node into a node in level 1 with degree value 1.

The word cloud can be replaced by a combination of n-grams. Searching via strings and phases are also possible within the backend algorithm. However, such searches are *NP-hard* and may increase the search time significantly.

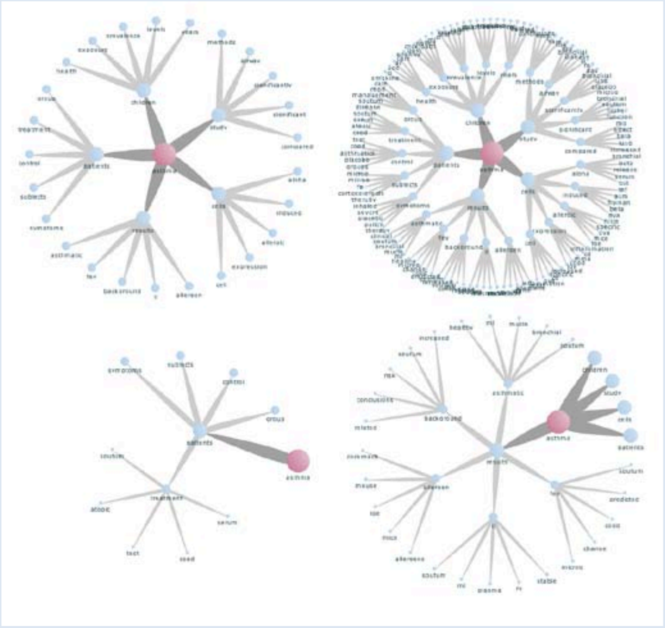


Figure 3. This figure shows the metamorphosed shapes of topology visualization in SEACOIN. Users can change the center of graph, degree of separation, and depth of information.

Interactive metamorphose topology visualization The visualization representation of the topological graph is generated using the open source visual analytics framework Birdeye Ravis[15]. The backend engines that drive the updates were developed in house. They include machine learning, text-mining, and topic modeling algorithms, and a specialized non-indexed database. These techniques include a general-purpose optimization-based discriminant analysis modeling framework and computational engine (DAMIP)[16-19], which has proven to be a powerful and flexible classification model and solution engine. Its multi-group prediction capability has been demonstrated successfully across a broad spectrum of biological and medical applications[20-24]. Working together, they generate in real-time multiple levels of hierarchy based on the most important words. These words are updated for every query. The node indicates a word resulting from analyzed abstracts of retrieved literature as shown in Figure 3. The center red node is the root of the search. The blue nodes represent child nodes. The size of the node represents the level in the hierarchy of words; a larger size indicates a higher-level word. Each node has a set of children except the words in the last level. We limit the maximum number of children (5 per node) and depth (4 levels) in our implementation in order to control information overload by filtering options and for ease of navigation within the visualization process. Advanced users can select the level of children and depth in their search to tailor to their needs.

Each edge connecting 'Parent node' to 'Child node' indicates that the associated word of the Child node is one of the top five most common words in abstracts filtered by the words of all ancestral nodes. Thus, the level 0 word is the most common word of all recalled abstracts. The five level 1 words, child nodes of a node at level 0, are the most common words in abstracts filtered by the word of level 0. For example, if users search on 'asthma', the most common word of the recalled abstract is 'asthma.' Users can see 'patients', 'study', 'results', 'cells', and 'children' in level 1 as shown in Figure 3 (top left). Each node in level 1 also has five children nodes, which are the most common words of abstracts filtered by 'its own ancestors' such as 'patients' AND 'asthma'. As users navigate to the lower level, they are able to retain more detail and specific words. For example, visualization search by keyword 'asthma' returns a number of formal/professional biological words such as ‘mRNA,’ a molecule of RNA encoding a chemical "blueprint" for a protein product and ‘COPD,’ Chronic Obstructive Pulmonary Disease in Level 3. Users can retrieve a concise list of literature by selecting the words in the lowest level.

The dynamic topology visualization facilitates summarization of numerous returned abstracts, and allows content-driven summary selected by users so as to avoid information overload. Users can control the graph's degree of separation as shown in Figure 3, and they can increase or decrease the depth of summarization. Also, users can manage the amount of information returned by filtering via a specific word selection. When they select a word in the visualization tree, SEACOIN will return only the literature that contains the word and its parent words. Moreover, users are able to change the center node of visualization and thus they can re-organize the summary based on the essential word they choose, and discard unwanted parts of returned information.

Table of recalled literature SEACOIN produces the list of recalled literature and presents them in a panel on the right side of the screen. The list is filtered and optimized as users interact with the word cloud and/or topology visualization. The articles are sorted by how deeply they relate to the selected words. When a user selects an article, its details are presented in a pop-up window. Also, users can extend their search to PubMed by clicking on the button that takes users to the article in PubMed.

SEACOIN system architecture (Figure 4) SEACOIN is developed via FLEX and Java. FLEX specializes in user interface while Java provides efficient technical options to manage large data sets. To accommodate unstructured information with full-text, a local database with Apache Lucene, a Java-based indexing and search library, is implemented. Apache Lucene enables considerably faster performance in full text searching[25] than MySQL. After Apache Lucene retrieves literature by matching keywords, Mallet[26], a Java-based package for statistical natural language processing, tokenizes the abstract of recalled literature, counts each word in the file, and returns top thirty common words. After initialization and selecting the main word for the root, the level 0, Apache Lucene matches not only the current word but also words of all the parents' word. After the process, Mallet returns the top thirty words again among the words in abstracts resulting from matching itself and its parents, to ensure that each of the words in the topology visualization tree is not overlapped. In-house topic modeling, text-mining, and machine learning techniques are also incorporated to speed up the analysis process. The process to find child nodes iterates until SEACOIN completes words from level 0 to level k. If we restrict the number of child nodes at each level to be n, then the total number of nodes generated at the end is n^k -1.

Result and Evaluation

The design of SEACOIN addresses expressed user concerns related to i) depth of retrieved information, ii) familiar and simple-to-use interface, iii) information overload, iv) number of returned pages, and v) result page format that is conducive to users' understanding of results, vi) relevance of filtered results and user needs; vii) representation of word cloud to query word entered by the user, and viii) time in performing the search. The depth of information control, the navigational power, the user control of selected nodes (topics) for exploration, and the 1-page visual-data summarization of SEACOIN address metrics i), iii), and iv). ii) and v) are user subjective. Anecdotal comments from bioinformatics researchers who have used SEACOIN have been favorable. In future research we will work to define formal metrics for evaluation. Herein, to shed some lights for metrics vi) – viii), we compare SEACOIN results with other search tools. We briefly describe the outcomes for searches of five common diseases (heart disease, cancer, Alzheimer's disease, diabetes, and asthma).

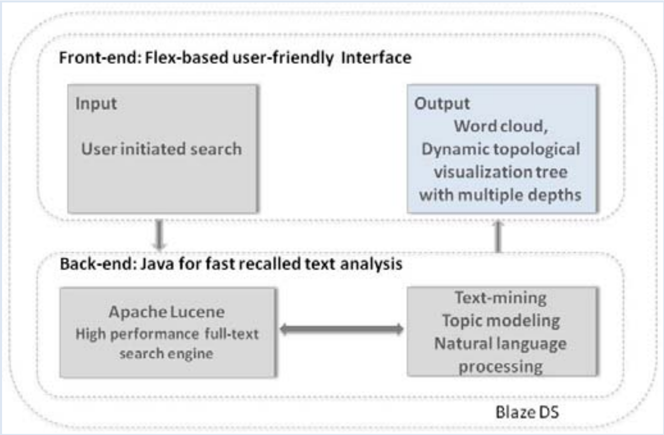


Figure 4. Systems architecture for SEACOIN: Frontend FLEX user interface, backend Java for fast recalled text analysis, with blazeDS combining frontend and backend

We search the five disease keywords in SEACOIN, Anne O'Tate[10], and 'Also try' of PubMed[27] that suggests alternative queries to filter and navigate originally retrieved results. SEACOIN is the only tool with visualization capability and process all the filtering in a page by selecting a node of topology visualization or a word from the word cloud. The other two are text-based search. Therefore, in order to utilize a suggested query, users need to open a new page. Anne O'Tate provides 'important words' resulting from analyzing abstracts of originally retrieved literature. When users select a word in the 'important word list', they can obtain new results by issuing a new query that is composed of the original query and a selected word among suggestions. PubMed deploys the 'Also try' feature that suggests possible keywords that might help users reach targeted literature. 'Also try' features statistical results from accumulated user logs by observing users' keyword queries. If users enter a word, 'Also try' suggests a new phrase composed of the original word and a new word. If users enter two words, 'Also try' suggests phrase containing the original two words and a new word. The filtering rates and time performance among the three approaches are summarized in Table 3.

Table 3. Filtering rates and time efficiency of SEACOIN, "Also Try", and "Anne O'tate"

	SEACOIN				"Also Try"				"Anne O'Tate"			
Search Key	lv. 1	lv. 2	lv. 3	Time (min)	lv. 1	lv. 2	lv. 3	Time (min)	lv. 1	lv. 2	lv. 3	Time (min)

Cancer	34%	12%	0.1%	3:30	9%	1%	0.7%	0:19	0.06%	0.01%	0.01%	5:40
Diabetes	35%	13%	0.3%	0:47	2%	1%	0.2%	0:19	0.06%	0.003%	0.001%	7:37
Heart	49%	14%	0.8%	0:13	25%	4%	0.5%	0:15	0.02%	0.001%	0.001%	5:09
Asthma	40%	18%	0.9%	0:13	28%	3%	N/A	0:14	21%	4%	0.5%	14:42
Alzheimer	47%	22%	1.5%	0:13	21%	22%	N/A	0:13	10%	10%	1.5%	2:30
Average % of literature remaining	41%	16%	0.7%		17%	6%	0.4%		6%	3%	0.04%	

We observe that Ann O’Tate reduces the largest number of articles in the first level with only 6% remaining at level 1. SEACOIN provides the most gradual filtering rate, and is the most consistent level by level across all 5 search keys. The filtering rates of Ann O’Tate and ’Also try’ vary depending on the search keyword. ’Also try’, which originated from accumulated user logs, lacks suggestions for keywords that are not searched often and stored in logs. For example, when ’asthma’ and ’Alzheimer’s disease’ are the search keys, only first and second level queries are suggested.

According to biomedical users who have tested the system, SEACOIN provides reasonable suggestions to drill down the recalled literature. The popularity of ‘Also try’ feature proves that the suggested words meet the users’ actual needs when filtering or navigating retrieved literature. SEACOIN covers more than 50% of ‘Also try’ suggestions through topology visualization. For example, when ’cancer’ is the search key, ‘Also try’ suggests ’breast cancer’, ‘lung cancer’, ‘prostate cancer’, ‘cancer stem’, and ‘colorectal cancer’. SEACOIN also suggests ‘breast’, ’lung’, and ’prostate’ as possible words to combine with ’cancer’. However, ’important words’ of Anne O’Tate rarely overlap with those of ’Also try’ of PubMed or SEACOIN.

We also note that the word cloud of SEACOIN contains words that are sufficiently representative of the query word entered by the user. All the words from the word cloud are similar to the word set obtained by applying the Topic Modeling technique directly to the same data set. Topic modeling in Mallet enables users to analyze large volumes of unlabeled text based on a probability model[26]. Using contextual clues, topic models can connect words with similar meanings and distinguish between uses of words with multiple meanings[26]. SEACOIN and Topic Modeling word cloud results share an average of 70% similarity among the five search keys. Specifically, 75% of the words with search key ’cancer’ are the same (Figure 5). Compared to the result of Topic Modeling, word cloud of SEACOIN obtained by tokenizing and counting frequency of each word in retrieved abstracts appears to be more appropriate because of its effort in avoiding word iteration. From these results, we can say that word cloud of SEACOIN provides a sufficient summary of the retrieved literature with meaningful words.

We observe some limitations in SEACOIN. Specifically, words at lower levels are composed of numerous abbreviations. For example, abbreviations such as TNF, mRNA, lgE, COPD, microg appear in the third and forth levels in SEACOIN. Since we standardize all the words into lowercase, recognizing the meaning of an abbreviation is tricky. For example, mRNA is written as mrna, which loses its obvious biological reference. Improvement has to be made to filter abbreviations and to ensure that case sensitivity is maintained.

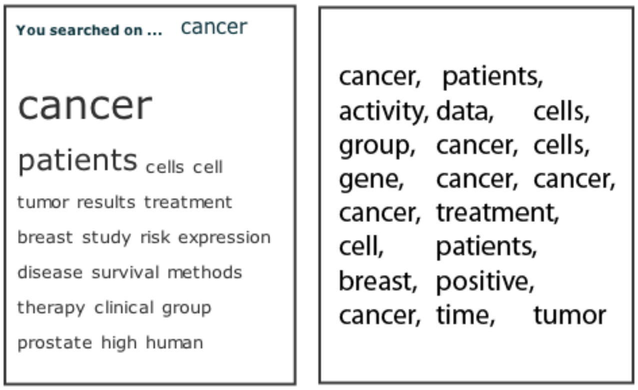


Figure 5. SEACOIN tag cloud versus Topic Modeling results of cancer. The two procedures share 75% common words.

Summary and Discussion

Considerable research has been conducted on technical solutions that assist biomedical researchers in literature search. In this work, we present **S**earch **E**xplore **A**nalyze **CO**nnect **IN**spire, SEACOIN, a web-based utility that attempts to alleviate the shortcomings of existing search technology. SEACOIN is designed for use by biomedical researchers whose time is limited and who can benefit from a simple-to-use visualization-based navigational search tool that can present the results of their queries in an easy-to-understand and content-driven summarized format. The system 1) performs real-time text mining, topic modeling, and natural language processing and development of user-driven hierarchy that focus on the users' needs; 2) enables users to navigate with interactive tools for topology visualization metamorphosis of concepts with multiple depths; 3) leverages a familiar word cloud visual interface to increase accessibility to the metamorphoses graph construct; and 4) avoids information overload by focusing on only essential functions with interactive visualization constructs that allow users to navigate freely and control the amount of information.

SEACOIN does not require special knowledge or experience in information retrieval technology. Users can utilize the system just as they use the Internet in everyday life. The system offers a user-friendly interface in FLEX. Specifically, the frontend FLEX user-interface is integrated with a backend Java system through BlazeDA that allows for fast analysis of recalled text. This design ensures that there is no compromise in system performance. FLEX facilitates the design of a user-friendly visualization interface that is pleasing to biomedical users. A word cloud, one of the most familiar visualization constructs and popular tools in social application sites, provides a quick overview of retrieved literature and allows users to focus their search results without the need to modify or re-input the initial query. Synchronization between the word cloud and an interactive visualization graph helps novice users to navigate their search. The system empowers users with the ability to control the depth of information that they desire to navigate (through the retrieved literature) without overwhelming them with voluminous data. Further, users can shift the focus of the words by clicking on a node of interest and exploring the selected depth of information there. This depth of information control is a distinct feature of SEACOIN. Although the confidence value of Alibaba[9] leverages the number of visible nodes, it decreases or increases the number of relationships without considering the depth of information. For example, when users search on cancer, SEACOIN enables them to control degree of separation about the topic cancer and Alibaba to control the number of relationships.

SEACOIN provides users with a 1-page visual-data summarization. When users select a node in the topology visualization tree, it filters the results set into literature where the word of a node and words of its parent node co-occur. Selecting a node of upper depth returns more results, while selecting a node at lower depth returns a subset of results. Anne O'Tate also filters the results when a user selects an entity resulting from parsing the retrieved literature. However, SEACOIN processes all filtering without changing a page, thus eliminating the burden of page formatting or re-ordering of pages. Same-page processing enables short performance time to filter results by loading every entity at once. It also offers a non-disruptive visual effect on the retrieved information.

SEACOIN generates different levels of information for a topic that users want to search in real-time basis to help users to retrieve in-depth information of retrieved literature. The real-time word hierarchy produced by parsing and analyzing the retrieved literature provides different levels of information focusing on the needs of the users. This is very different from a pre-defined ontology where users are conformed to the preset format and thus limit their active participation and generation of useful data that best fit their needs.

SEACOIN's strength is in analyzing text-based data, the real-time generation of the content relationship, and the resulting navigational capability. We must ensure its scalability and extensibility. Query times for the five instances presented herein range from seconds to a few CPU minutes. Future work will involve defining formal metrics for evaluation, benchmarking the computational effort for other searches, including searches of other scientific literature databases besides PubMed. We must improve the handling of abbreviations so that important biological/clinical meanings of terms are not lost. More generally, user accessibility and systems design feedback will be sought and incorporated. The system has also been tested for understanding connectivity among authors, institutes, topics, funding sources, etc., and we expect continued advances on this front.

Acknowledgement

This work is partially supported by grants from the National Science Foundation, and by the PHS Grant UL1 RR025008 from the Clinical and Translational Science Award program, National Institutes of Health, National Center for Research Resources.

References

1. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet 2006;7:119-29.
2. Herskovic JR, Tanaka LY, Hersh W, Bernstama EV. A day in the life of PubMed: Analysis of a typical day's query log. J Am Med Inform Assoc 2007;14:212-20.
3. Dogan RI, Murray GC, Névéol A, Lu Z. Understanding PubMed user search behavior through log analysis. Database. 2009:bap018.
4. Davies K, Harrison J. The information-seeking behaviour of doctors: a review of the evidence. Health Info Libr J 2010;27:341.
5. Douglas SM, Montelione GT, Gerstein M. PubNet: a flexible system for visualizing literature derived networks. Genome Biol 2005;6:R80.
6. Perez-Iratxeta C, Bork P, Andrade MA. XplorMed: a tool for exploring MEDLINE abstracts. Trends Biochem Sci 2001;26:573-5.
7. Doms A, Schroeder M. GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res 2005;33:W783-6.
8. Eaton AD. HubMed: a web-based biomedical literature search interface. Nucleic Acids Res 2006;34:W745-7.
9. Plikus MV, Zhang Z, Chuong CM. PubFocus: semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithm. BMC Bioinformatics 2006;7:424.

10. Smalheiser NR, Zhou W, Torvik VI. Anne O’Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results. J Biomed Discov Collab 2008;3:2.
11. Xuan WJ, Dai MH, Mirel B, Song J, Athey B, Watson SJ, Meng F. Open Biomedical Ontology-based Medline exploration. BMC Bioinformatics 2009;10:S6.
12. Hunter L, Cohen KB. Biomedical language processing: what’s beyond PubMed? Mol Cell 2006;21:589-94.
13. D’Amico G, Del Bimbo A, Meoni M. Interactive visual representations of complex information structures. Lect Notes Comput Sc 2010;91:101-12.
14. Rivadeneira AW, Gruen DM, Muller MJ. Getting our head in the clouds: toward evaluation studies of tagclouds. SIGCHI 2007;1-2:995-8.
15. Bertini M, D'Amico G, Ferracani A, Meoni M, Serra G. Web-based semantic browsing of video collections using multimedia ontologies. ACM Multimedia 2010:1629-32.
16. Gallagher RJ, Lee EK, Patterson DA. An optimization model for constrained discriminant analysis and numerical experiments with iris, thyroid, and heart disease datasets. Proc AMIA Annu Fall Symp. 1996:209-213.
17. Lee EK, Gallagher, RJ, Patterson, D. A linear programming approach to discriminant analysis with a reserved judgment region. . INFORMS Journal on Computing. 2003;15:23-41.
18. Lee EK. Large-scale optimization-based classification models in medicine and biology. Ann Biomed Eng. 2007;35:1095-1109.
19. Brooks JP, Lee, EK. Analysis of the consistency of a mixed integer programming-based multi-category constrained discriminant model. Ann Operat Res Data Mining. 2010;174:147-168.
20. Lee EK. Machine learning framework for classification in medicine and biology. Integration of artificial intelligence and operations research techniques in constraint programming for combinatorial optimization problems. CPAIOR 2009. 2009;5547:1-7.
21. Lee EK, Wu, TL. Classification and disease prediction via mathematical programming. In Data Mining, Systems Analysis, and Optimization in Biomedicine. AIP Conference Proceedings. 2007;953:1-42.
22. McCabe MT, Lee EK, Vertino PM. A multifactorial signature of DNA sequence and polycomb binding predicts aberrant cpg island methylation. Cancer Res. 2009;69:282-291.
23. Querec TD, Akondy RS, Lee EK, et al. Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. Nat Immunol. 2009;10:116-125.
24. Nakaya HI, Wrammert J, Lee EK, et al Systems Biology of Vaccination for Seasonal Influenza in Humans Nature Immunology 12, 786-795, 2011.
25. Palidwor GA, Andrade-Navarro MA. MLTrends: Graphing MEDLINE term usage over time. J Biomed Discov Collab 2010;5:1-6.
26. McCallum AK. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu/. 2002.
27. Lu Z, Wilbur WJ, McEntyre JR, Iskhakov A, Szilagyi L. Finding query suggestions for PubMed. AMIA Annu Symp Proc 2009:396-400