INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 76 (2007) 780–789

journal homepage: www.intl.elsevierhealth.com/journals/ijmi

An end user evaluation of query formulation and results review tools in three medical meta-search engines

Gondy Leroy[a,*], *Jennifer Xu*[b], *Wingyan Chung*[c], *Shauna Eggers*[d], *Hsinchun Chen*[d]

[a] *School of Information Systems and Technology, Claremont Graduate University, 130 E. Ninth Street, Claremont, CA 91711, United States*
[b] *Department of Computer Information Systems, Bentley College, United States*
[c] *Department of Information and Decision Sciences, The University of Texas at El Paso, United States*
[d] *Management Information Systems, The University of Arizona, United States*

ARTICLE INFO

Article history:
Received 30 March 2006
Received in revised form
17 July 2006
Accepted 7 August 2006

Keywords:
Information storage and retrieval
Unified Medical Language System
Informatics

ABSTRACT

Purpose: Retrieving sufficient relevant information online is difficult for many people because they use too few keywords to search and search engines do not provide many support tools. To further complicate the search, users often ignore support tools when available. Our goal is to evaluate in a realistic setting when users use support tools and how they perceive these tools.

Methods: We compared three medical search engines with support tools that require more or less effort from users to form a query and evaluate results. We carried out an end user study with 23 users who were asked to find information, i.e., subtopics and supporting abstracts, for a given theme. We used a balanced within-subjects design and report on the effectiveness, efficiency and usability of the support tools from the end user perspective.

Conclusions: We found significant differences in efficiency but did not find significant differences in effectiveness between the three search engines. Dynamic user support tools requiring less effort led to higher efficiency. Fewer searches were needed and more documents were found per search when both query reformulation and result review tools dynamically adjust to the user query. The query reformulation tool that provided a long list of keywords, dynamically adjusted to the user query, was used most often and led to more subtopics. As hypothesized, the dynamic result review tools were used more often and led to more subtopics than static ones. These results were corroborated by the usability questionnaires, which showed that support tools that dynamically optimize output were preferred.

1. Introduction

During the last few years, attention in medical informatics has shifted away from information retrieval (IR). However, researchers still need to search the literature and they are still using simple search engine techniques. Most existing IR research evaluates the performance of search tools and measure precision and recall with laboratory experiments. For example, Baujard et al. [1] evaluated their multi-agent retrieval software in terms of precision and recall of web pages for pre-specified medical queries. Bin and Lun [2] compared the retrieval effectiveness of eight medical online search engines with single keyword and question-answering tasks. For an historic overview of the usage of laboratory studies, we refer to

* *Corresponding author*. Tel.: +1 909 607 3270; fax: +1 909 621 8564.
E-mail address: gondy.leroy@cgu.edu (G. Leroy).
1386-5056/$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2006.08.001

Su [3]. Such studies shed light on individual components and provide estimates of optimal results. However, they often rely on gold standards and simulations instead of end users to evaluate the tools and they are seldom executed in a realistic setting. Fortunately, we may be seeing a shift towards an end user perspective. For example, Gaudinat et al. [4] provide a brief indication of how informative, reliable and trustworthy end users believed their search engine to be.

We believe that it is important to study how end users interact with and evaluate their experience with search engines. They will be the ultimate users and, even though they might believe some document to be relevant when an expert would disagree, their experience with the system will determine whether or not they will decide to use it. We see our study as complementary but essential to laboratory evaluations of specific components. This is especially the case with IR because contradictions are found between efficiency and usability of tools and the actual usage, or lack thereof in reality.

In the study presented here, we focus on tools that help users construct queries and evaluate results. Three meta-search engines that incorporate comparable support tools are compared. We developed two of the meta-search engines, MedTextus and HelpfulMed, and compare them with a third, the National Library of Medicine's (NLM) Gateway.

2. Theoretical foundation

Sutcliffe and Ennis [5] provide a framework to evaluate information retrieval systems that extends earlier work by both Kuhlthau [6] and Marchionini [7]. The framework comprises four steps during which different activities and strategies are executed. During *problem identification*, a user learns that she needs to retrieve information. During *need articulation*, she can consult external sources such as controlled vocabularies or use her own domain knowledge to formulate terms to search. This leads to *query formulation* or combining terms in a query suitable for the specific information retrieval system. With common search engines, these second and third steps are identical and consist of providing a keyword list. However, large differences between these steps exist when users can form precise, complex queries that include, for example, Boolean terms or field descriptors. The efficiency of the query formulation will then also depend on the user's device knowledge. Finally, during *results evaluation* the information need and the information retrieved are compared. Our works address the third and fourth steps of the framework: query formulation and results evaluation.

2.1. *Query formulation*

Although having sufficient search terms to find the few relevant documents among millions is vital, users use only about two keywords when searching the Internet [8–11]. Lau and Horvitz [12] looked at different search topics for variations in the number of keywords and found that health-related queries did not differ from this average. In response to this problem, query expansion, manual or automated, can be used. With manual, i.e., interactive query expansion, users themselves indicate which terms should be used for expansion. With automated expansion, a system selects the terms. Both automatic and interactive query expansion have been studied at the Text REtrieval Conferences (TREC) and Hawking and Craswell [13] concluded that, in general, better results are obtained when some form of query expansion is used. In medicine, Hersh et al. [14] tested different Unified Medical Language System (UMLS) Metathesaurus components and found that the results improved for some queries but only with synonyms. French et al. [15] simulated queries and expanded them with Medical Subject Headings (MeSH) to successfully improve retrieval performance.

Although automatic query expansion may be easier for users, it may also lead to a feeling of losing control. This explains why users often prefer manual query expansion [16]; it is also the reason why we chose manual query expansion for our approach. With interactive query formulation, the presentation of the terms and the experience of the users matter. For example, Joho et al. [17] compared a hierarchical and a list presentation of terms. Retrieval performance itself was not affected, but users needed less time to form a query with the hierarchically presented terms. Unfortunately, users seldom use query expansion tools spontaneously [18,19]. Jansen et al. [20] found that users requested query expansion only 5% of the time or less. McCray and Tse [21] found that when the system suggested a correct alternative for a misspelled term users used this alternative in only 45% of the cases. Moreover, users are not proficient at selecting good terms. Ruthven [22] found that subjects vary widely in their ability to select good expansion terms or identify poor expansion terms.

2.2. *Results evaluation*

The most common output format of a search engine is a ranked list of documents. The exact parameters of the ranking are often unknown and users are required to browse through the list. Previews or overviews may expedite results review [23,24]. A preview is extracted from the original document and acts as a surrogate. It is effective when it communicates sufficient information to the user about the content. For example, most search engines provide an excerpt of text, called a snippet. Snippets are short previews of the information that can be found in the document. Some simply show the first few lines of a text, others display the text surrounding user keywords, some use heuristics to select sections of documents [25] and more advanced previews display a summary of the document. Previews can be textual, graphical or a combination of both. Woodruff et al. [26] compared textual and graphical (thumbnails) summaries and found the best results when combining both types. In contrast to a preview, which is based on a single document, an overview is based on a collection of objects and is effective when it provides an immediate understanding of the size, extent and content of the collection. Shneiderman [27] provides a taxonomy of overview methods that includes one-, two- and multi-dimensional methods, temporal structures and networks.

Few previews and overviews have been tested in medicine. Pratt and Fagan [28] compared dynamic categorization of search results with common relevance ranking and clustering. The documents retrieved for a user query were divided into categories based on the words in the query and their

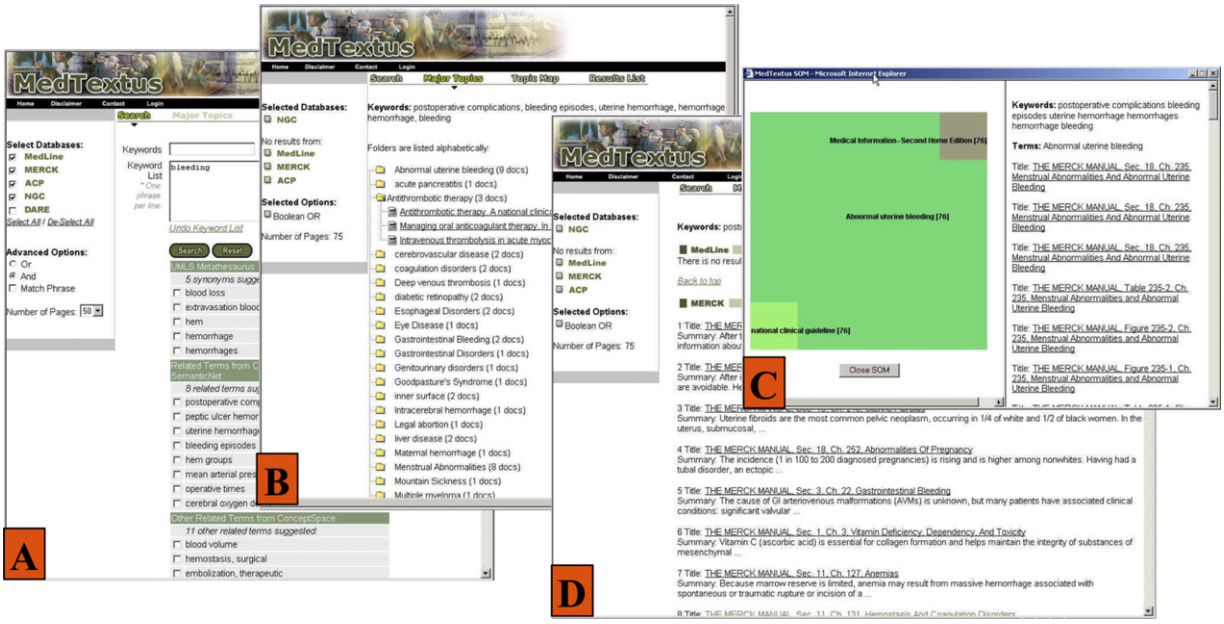

Fig. 1 – MedTextus interface. A: keyword suggestion; B: major topics folders; C: major topics map; D: result list.

association with the UMLS Metathesaurus and MeSH vocabularies. Patients could find information faster and were more satisfied with this approach. Another example is Proteus-BIO, a search engine for disease outbreak reports [29]. The information about disease outbreaks, time, location and victims is extracted from the documents. Users search with keywords and are presented a tabular format with relationships between the extracted elements that match their keywords. Each relation acts as a preview of the content of the associated documents and allows users to find more relevant documents in a shorter time period. A similar approach for biomedical text is used in Genescene [30]. Here, an interactive graph visualization of biomedical relationships is added to a tabular representation. Beier and Tesche [31] used a simpler approach and provided an overview of a meta-search based on the origin of the results, such as Internet sites or journals.

3. Research questions

We are interested in finding out if support tools requiring more or less user effort to use and understand will lead to different levels of spontaneous use, effectiveness and efficiency. For this purpose, we developed two meta-search engines, HelpfulMed and MedTextus, and compared them with a third existing one, NLM Gateway.

Sutcliffe and Ennis [5] predict that users will use query formulation tools to help form their search query when their domain knowledge is low. Others have indicated that users do not use the tools [19–21] and do not expend a lot of effort searching [5,32,33]. We hypothesize that tools will be used more readily and will be preferred when they require less effort from the user. Furthermore, the permissiveness of the overview tools [34] influences acceptance and use. A permissive interface is an interface that allows different paths to the same goal, leading to higher acceptance and better results. Finally, we expect that more dynamically adjusted support tools will allow users to be more effective and efficient in finding relevant information.

4. System descriptions

4.1. *MedTextus*

MedTextus (Fig. 1) is a meta-search engine of five medical databases: MEDLINE, the Merck Manual, the Database of Abstracts of Review of Effectiveness (DARE), the National Guideline Clearinghouse (NGC) and the American College of Physicians' database (ACP). Users search with MedTextus by typing their terms in a keyword box: one phrase per line. The search terms can be combined with a Boolean operator. Users also choose which databases to search and how many documents (25, 50, or 75) to retrieve per database.

To help formulate a query, users can consult the keyword suggester, which provides three sets of keywords dynamically selected and adjusted for the user query. The first set of terms consists of synonyms for each user keyword retrieved from the UMLS Metathesaurus. The second set provides new, non-synonymous keywords based on a concept mapping algorithm developed earlier [35,36] and tested for the first time in a realistic setting. The third set of keywords is retrieved directly from the thesaurus and not filtered with the Semantic Network.

MedTextus's interface is very permissive and shows the results in three different manners. It includes two overview tools and the classical results listing. The overview tools organize the results in categories, which is recommended as a best practice by Resnick and Vaughan [18]. The topic folders dynamically combine the results from the different databases. Search results are preprocessed to select only relevant text, e.g., not the copyright notice, and extract all medically relevant noun phrases. Folder labels are chosen based on a set of

Fig. 2 – HelpfulMed interface. A: keyword suggestion; B: major topics map; C: result list.

heuristics, developed during pilot tests. First, we extract a set of candidate labels based on all noun phrases in the results. Noun phrases that contain four words or more, that appear in our stop term list, or that contain a word from our stop term list are excluded. Singular and plural noun phrases are represented only once. From these candidates we select the folder labels. A phrase that appears multiple times in the list becomes a folder label. This provides a natural link between two documents. A phrase that appears only once becomes a folder label when it contains a user search term. Users are especially interested in the documents that contain their terms. In addition, a phrase that appears only once becomes a label when it is part of the UMLS Metathesaurus. All folders are shown, ordered alphabetically and displayed with the number of documents they contain. Documents can appear in multiple folders. Showing only a subset of folders (e.g., those with frequently recurring terms or with multiple documents) raised questions during pilot studies. Clicking on a folder displays the documents contained in that folder and their link to the original database. The folder view is shown automatically to the users after submitting a search to the search engines. Users can, however, switch immediately to the other views.

The topic map is an abstract two-dimensional representation of the results. It dynamically categorizes documents with the self-organizing map (SOM) algorithm [41], a neural network approach that places documents in a cell on a grid. Documents within a cell are similar to each other but dissimilar to documents not in that cell. In MedTextus, a unique map is constructed on-the-fly for each search result. When users click on a region of the map the documents belonging to that region are shown, and users can follow the links back to the document in the original database. Finally, a result list that shows all results is also available. The rankings from the underlying databases are retained and all documents are linked to the original database.

4.2. *HelpfulMed*

HelpfulMed (Fig. 2) provides access to the same five medical databases as MedTextus. It has one search box where the keywords can be typed. It also uses the same automatically created thesaurus as MedTextus but shows more terms. Two columns with terms ordered according to the thesaurus co-occurrence score are provided: the first shows all terms and the second shows the subset consisting of MeSH terms. Terms extracted as noun phrases by the AZ Noun Phraser are tagged with an “N”, terms that are part of MeSH are tagged with an additional “M”. Each term also has a letter indicating to which keyword it is related. In addition to phrases, HelpfulMed presents author names based on co-occurrence of the author name with the search term. Users can check in a checkbox which terms they want to include in their query.

Users get a results list, which retains the ranking per source, comparable to the result list in MedTextus. However, they can also browse an additional map displaying 10 million MEDLINE abstracts. This map is also based on the SOM, described above, but it is static and is not specific to individual user queries. There is an alphabetic folder list of the categories adjacent to the graphical display. Its usage is similar to the dynamic map in MedTextus.

4.3. *NLM Gateway*

NLM Gateway is a medical meta-search engine intended to be used by Internet users who are unfamiliar with NLM’s resources [37]. In addition to the PubMed/MEDLINE databases,

Table 1 – Topics

Please discuss cell-mediated immune defense
Please discuss split genes
Please discuss organ and cell transplantation
Please discuss retroviral oncogenes
Please discuss growth factors
Please discuss cholesterol metabolism
Please discuss restriction enzymes
Please discuss the new recommendations for unexplained infertility
Please discuss the provision of diagnostic virological services to a district hospital
Please discuss the vaccine strategies that can be employed for the prevention of common childhood virus diseases
Please discuss the quality control of serological assays
Please discuss the monitoring of virus infections in pregnancy

NLM Gateway provides access to books, serials, audiovisuals, computer files, meeting abstracts and health service projects. Users can search the underlying information sources with search-by-field features such as subject search (by default), author search and title search. Boolean operators are also available, as well as options to limit the results by language and publication date.

NLM Gateway uses MeSH for keyword suggestion. A keyword is mapped to its corresponding MeSH terms. The definition and MeSH trees for these terms are also provided. Users browse this information for additional terms and add them to their query by checking the box, choosing the Boolean operator and clicking the "Add to Search" button. NLM Gateway provides a simple result review feature. Search results are organized in a table that lists the number of documents found for each type of resource, e.g., journals or books. A user clicks on the "Display Results" button to view the results. After selecting the category of results, a typical list is shown.

5. Methods

5.1. *Design*

The three meta-search engines described provide comparable query formulation and results review tools. However, each requires more or less user effort. The query formulation tools in MedTextus require the least effort and provide a short list of terms customized for the user's query. HelpfulMed provides a much longer list of terms, which is still optimized for the query as a whole. NLM Gateway requires the most effort; it provides a long list of terms for each user keyword and does not optimize for the query as a whole. MedTextus also provides the most permissible overview of the results; it contains three views of the results: folder, map and listing. HelpfulMed and NLM Gateway provide a basic listing of the results. We expect that higher permissiveness will allow more people to use the meta-search engine effectively, which will be reflected in a higher usability score.

We used a within-subjects design for our study so that each user worked with all three meta-search engines. The order in which the search engines were presented was varied with a balanced approach to avoid ordering effects. Twelve topics (Table 1) to be used as tasks were randomly assigned to the users and meta-search engines.

5.2. *Procedure and tasks*

We recruited medical students, professionals and librarians to participate in our study. Users were first asked demographic and background questions. All questions were framed as positive statements, e.g., "I am an expert online searcher", and users indicated on a seven-point Likert scale their agreement or disagreement with the statement. There was also a "Not Applicable" option. To ensure that all users needed to consult the literature before completing the task, we opted for advanced topics (Table 1). In the framework of Sutcliffe and Ennis [5], this means that the users' domain knowledge was low and they would benefit from tools that help formulate queries and evaluate results.

The study was conducted with each user individually. A facilitator first explained the first meta-search engine interface and showed the user its functionality during a practice session using the keyword "cancer". In this manner, all users, regardless of their expertise, gained familiarity with all components. In the framework of Sutcliffe and Ennis [5], this means that device knowledge was the same for all users. Users were then provided with the first task. We used a browse task to guide the user interaction. We asked users to imagine they had to write a research paper, a task they were all familiar with. They were given the theme and asked to find documents that provide alternative opinions, discuss the topic entirely or focus on a relevant subset of information. They were asked to record the subtopics and supporting abstracts (abstract id) for each subtopic. We asked users not to read entire documents but to scan them and indicate which ones they would consider for further reading, e.g., which ones they would print. This was done so that an experimental session could be concluded in 1 h. When the user indicated they had found sufficient information, the second and later third, meta-search engines were introduced and the scenario was repeated.

5.3. *Measurements*

We calculate both effectiveness and efficiency based on the subtopics and supporting documents written down by the user. The number of user-selected subtopics and the average number of abstracts per subtopic indicate the effectiveness of the meta-search engine. The average number of searches and the number of user-selected documents per search indicate the efficiency of the meta-search engine. We accept the

Table 2 – Background questions (N = 23)

	Average score
Computer and search expertise	
I am an expert online searcher	3.0
I search online for information daily	2.3
I enjoy trying out new software	2.7
I enjoy playing computer games	4.6
I am an expert user of NLM Gateway	4.8
Controlled vocabulary expertise	
I am an expert using MeSH terms	4.4
I am an expert using ICD terms	6.1
I am an expert using ------ terms	2.0
User task topics	
I know the answer to the following question: topic 1 inserted	3.8
I know the answer to the following question: topic 2 inserted	3.7
I know the answer to the following question: topic 3 inserted	4.1

Likert scale 1–7, a smaller number represents more agreement with the statement.

subtopics and supporting documents at face value because of our end user perspective: we do not evaluate the correctness of each underlying medical database nor the actions of the end users but the usefulness of the interfaces for users.

At the end of a session with each meta-search engine, users filled out the usability questionnaire, which contained four sections. The first section had one question, which asked users for their overall opinion of the interface. The next section consisted of a general usability questionnaire developed by Lewis [38] with 19 questions that represent three characteristics: system usefulness, information quality and interface quality. The third section contained seven specific questions about the query formulation components. For each component, we asked if it was useful, relevant and easy to use. The last section contained five similar questions about the result evaluation tools. Similar to the background questions, we used a seven-point Likert scale.

At the end of the entire session, when users had worked with all three meta-search engines, we asked them to compare the meta-search engines and indicate their preference for each with respect to seven potential information tasks.

6. Results

Twenty-three users, 12 female and 11 male, participated in our user study. They were medical students, professionals and medical librarians. All had completed a bachelor's degree and 10 had already obtained a graduate degree. Those without a graduate degree were pursuing it at the time of our study. On average, the users considered themselves somewhat experienced online searchers (Table 2). Most did not have experience with NLM Gateway or controlled vocabularies, with the exception of two users who knew how to use NLM Gateway and two who considered themselves experts using MeSH.

6.1. *Effectiveness and efficiency*

For each user, the number of subtopics, the number of abstracts and the number of searches was counted. One "search" constitutes all query formulation activities until the user clicks the search button and submits a query to the meta-search engine. It also includes all evaluation activities with the results set of that particular submitted query. Table 3 provides an overview of the results. We performed ANOVAs with the meta-search engine as the independent variable (repeated measures).

There was a significant effect of meta-search engine for the number of searches performed, $F(2, 44) = 3.264$, $p < .05$, and also for the documents selected per search, $F(92, 44) = 7.716$, $p < .01$. It took users fewer searches with MedTextus than with HelpfulMed or NLM Gateway to find the subtopics. The number of documents found per search was higher with MedTextus (3.4 docs/search) than with HelpfulMed (1.3 docs/search) or NLM Gateway (2.1 docs/search). Post hoc comparisons showed that

Table 3 – Effectiveness and efficiency (N = 23)

	MedTextus	HelpfulMed	NLM Gateway
Effectiveness			
Average subtopics selected	3.8	3.5	3.4
Documents selected per subtopic	1.7	1.0	1.0
Efficiency			
Number of searches*	2.7	3.1	3.6
Documents selected per search**	3.4	1.3	2.1

*Significant effect at $p < .05$; **significant effect at $p < .01$.

Table 4 – Usage of components (N = 23)

% of users who used	MedTextus	HelpfulMed	NLM Gateway
Query expansion tools	52	74	52
Major topics categorization	100	–	–
Map categorization	74	74	–
Results list	52	96	100

Table 5 – Percentage of topics found with different components

Percentage of topics found	MedTextus	HelpfulMed	NLM Gateway
With query expansion	10	21	5
Without query expansion	90	79	95
Total	100	100	100
With major topics categorization	65	–	–
With map categorization	23	11	–
With results list	12***	89***	100
Total	100	100	100

***Significant difference, paired *t*-test, $p < .001$.

this difference between HelpfulMed and MedTextus was significant (Bonferroni adjustment used, $p < .05$). Although users found the most subtopics with and also retrieved on average more than one abstract per subtopic with MedTextus, these differences were not statistically significant.

To ensure that the results were due to the different interfaces and not the particular theme, we performed an outlier analysis for each of the 12 themes for each metric to show that there are no values that are extremely different, for example because the theme is extremely difficult or easy. To this end, we calculated *z*-scores for the number of subtopics reported by users, the number of searches performed, the number of documents reported and the number of documents per search. No score was so extreme as to be considered an outlier.

6.2. *Usage of query expansion and overview tools*

Query expansion was used most often with HelpfulMed, by 74% of the users, and by half of the users with MedTextus and NLM Gateway (Table 4). All users used the Topic Folders categorization in MedTextus. Seventy-four percent of users consulted the maps in MedTextus and HelpfulMed. The results list was used by fewer users in MedTextus (52%) than in HelpfulMed (96%). The use of the NLM Gateway list was mandatory (constant value).

The top section of Table 5 shows how often answers were found with the help of query expansion tools. On average, query expansion tools led to 5% of the topics with NLM Gateway, 10% with MedTextus and 21% with HelpfulMed. The bot-

Table 6 – Usability results (N = 23)

	MedTextus	HelpfulMed	NLM Gateway
Overall liking	2.8	3.5	3.3
Lewis' scale			
Total	2.6	3.3	3.3
Subscale: system usefulness	2.5	3.3	3.2
Subscale: information quality+	2.7	3.4	3.4
Subscale: interface quality	2.5	3.1	3.2
Query formulation components			
Total*	2.4	3.2	3.2
Subscale: synonyms	2.1	3.0	3.2
Subscale: related terms	2.0	3.0	3.3
Subscale: modifiers (and, or)	2.3	3.2	3.1
Results review components			
Total	2.8	3.6	3.5
Subscale: listing results	2.4	2.9	3.4
Subscale: folders	2.2	2.9	3.0
Subscale: map	4.1	4.3	

Likert scale 1–7, a smaller number represents a better score. +Trend at $p < .1$; *significant effect at $p < .05$.

Table 7 – Average ranking of the meta-search engines (N = 23)

Questions	Meta-search engine		
	MedTextus	HelpfulMed	NLM
To get one document, I would use	2.2	2.8	**1.7**
To get a set of documents, I would use	**1.9**	2.4	2.2
To find a specific answer to a question, I would use	**2.1**	2.3	2.5
To get an overview of answers to a question, I would use	**2.0**	2.3	2.1
When I do not know much about the subject, I would use	**2.1**	2.4	2.1
When I do not know many good keywords, I would use	**1.8**	2.0	2.4
When I need an exhaustive overview, I would use	2.0	2.5	**1.7**

Rankings 1–3. Top rank in bold.

tom section of Table 5 looks at the usage of the result review tools. In MedTextus, most of the topics (65%) were found when users used the Topic Folders. The dynamic map in MedTextus led to 23% of the topics, which is more than the percentage of topics found with the static map in HelpfulMed (11%). The results list led to significantly fewer results in MedTextus (12%) than in HelpfulMed (89%). In NLM Gateway all results were based on a results list.

6.3. *Usability*

Table 6 provides an overview of the usability results. Statements in the questionnaire were in positive format, e.g., "It was simple to use this system". Smaller numbers indicated a stronger agreement with the statement. The usability evaluations for HelpfulMed and NLM Gateway were similar. Overall, MedTextus received the best usability scores. An ANOVA indicated a strong trend for the information quality ($p < .1$) and a significant effect for the query formulation component, $F(2, 44) = 3.825$, $p < .05$, with the difference between MedTextus and HelpfulMed having the most impact.

As described above, we expected that users with different backgrounds might prefer different components. We therefore correlated each expertise-related question (Table 2) with usage of the meta-search engines (effectiveness and efficiency) and also usability using the Pearson product moment correlation coefficient. Two significant correlations were found: users who were frequent searchers ("I search online daily") liked NLM Gateway better ($r = .417$, $p < .05$) and also liked the synonym component of MedTextus better ($r = .527$, $p < .05$).

6.4. *Qualitative feedback*

At the end of the sessions with the three meta-search engines, we asked users to compare and rank all three with respect to seven tasks (Table 7). A score of "one" indicates the most preferred interface, "three" the least preferred. NLM Gateway and MedTextus were the most preferred meta-search engines. Most users preferred the query expansion tools and review tools in MedTextus. Several users commented on the maps. Most users seemed to either like or dislike the maps. To verify if this was a general trend, we looked at the usability evaluation of the maps in MedTextus and HelpfulMed and found a significantly positive correlation ($r = .486$, $p < .05$).

7. Discussion

We hypothesized that the usability evaluation and usage of the query reformulation tools would be the highest for MedTextus, which requires the least effort. The MedTextus query formulation tools were indeed considered most usable but usage of query formulation tools was highest for HelpfulMed. HelpfulMed provides long lists of potentially useful keywords dynamically optimized for the global user query compared to the short, dynamically adjusted list in MedTextus and the long, unadjusted lists in NLM Gateway.

The usage of query expansion tools resulted in more results for the two meta-search engines with dynamic query formulation support. HelpfulMed, which had the highest usage, also had the highest number of results based on these queries. In contrast, NLM Gateway, with query formulation usage similar to MedTextus, had the smallest number of results based on these modified queries. It may be that users are incapable of selecting good terms for expansion and that this is especially hard when the suggested terms are not optimized for the query as a whole. Such a lack of background knowledge would match results found by Ruthven [22] where subjects varied widely in their ability to recognize both good (30–75% recognition of good terms) and poor terms for expansions.

We also hypothesized that permissive result review tools would lead to higher usability evaluation and usage. MedTextus had the highest permissibility and its folder overview was used most often, leading to a higher percentage of results. This may be partially due to the fact that it was automatically shown to users. However, users were shown in advance how they could see the results in different formats, so it was not a lack of knowledge that led to this behavior. The usability scores confirm the preference for the folders: they received the highest usability ratings among all tested. In contrast, users considered the two map overviews either usable or unusable and these usability scores are the only ones that did not receive a better than average score. We conclude that users again opt for a dynamic solution. However, they prefer a representation they are familiar with, such as a list of folders, not a map.

We expected that because of the dynamic nature of all tools in MedTextus, it would lead in general to more effective and efficient searching. We found no significant differ-

Table 8 – Summary of lessons learned

Literature	This study
Query formulation	
Lay people seldom use query formulation tools [18–21]	Researchers and other domain experts use query formulation tools
Some type of query expansion is beneficial [13]	More query expansion does not lead to more results with those queries, it depends on the type of expansion tools available
Presentation of terms matter: users need less time to form a query with hierarchically presented terms [17]	Dynamically optimized query terms were preferred, used more often and led to more results
Results review	
Dynamic categorization is preferred by patients [28]	Dynamic categorization is preferred by researchers
List of results are most commonly available with current search engines	Users use and prefer categorized results when available

ences for effectiveness, although the numbers display the hypothesized direction. However, we found that users were more efficient when using MedTextus compared to NLM Gateway.

When ranking the three meta-search engines for different tasks, MedTextus was usually preferred. Users would prefer NLM Gateway for two tasks: to find one document or to provide an exhaustive overview. We believe this may be partially due to the authoritative status that NLM Gateway enjoys. In the case of one document, users may prefer to cite a well-known source. As reported by Fogg [39,40], a real world presence and reputation are factors taken into account by users. For an exhaustive overview, previews and overviews may not matter since users expect to read all documents. In contrast, MedTextus was generally preferred for the other tasks. We believe this is due to the previews/overviews that speed up the search process, e.g., when you need the answer to a specific question.

Table 8 contains a summary of the lessons learned from this study in comparison to current literature. We found that our users, who are not laymen, used the query formulation tools frequently although they were not explicitly asked to do this. Although some type of query expansion is considered beneficial, it did not necessarily lead to more results in our study. Furthermore, researchers also like dynamic categorization of the results. Tools that provide support dynamically optimized for the entire query and result sets were preferred and led to more results.

Our study has limitations that need to be taken into account. We did not look at individual keywords. An in-depth analysis may show interesting interactions between the types of keywords used, the resulting query expansion outcome and the user evaluations. In addition, users based their opinion on a single interaction with a meta-search engine. Training users for an extended amount of time may lead to different results and a clearer indication of the contribution of each support tool to the results. Finally, we indicated that more user-selected documents indicate a better result. This is correct from the end user evaluation standpoint because for users it was easier to find more documents of interest. However, from the viewpoint of algorithm evaluation, an expert would have to evaluate if all retrieved documents were relevant. We did not include that in our study, since the three meta-search engines access largely the same databases.

8. Conclusion

The purpose of our study was to provide an end user evaluation of query formulation and result review tools in a realistic setting. We tested three meta-search engines with different user support tools that access by and large the same databases. Each meta-search engine provided user support in a different fashion. The main difference for the query formulation tools lies in how tailored the information was to the user queries, resulting in more or less effort required by the user. Different search engines also provide multiple or a single view of the results. The users' overall liking was the highest for the meta-search engine that customized the output of its support tools to their queries. Providing a good structure, even if there are many results, is especially important. This was reflected in the evaluation of the individual components.

Acknowledgements

The authors would like to thank the development teams of the Artificial Intelligence Lab at the University of Arizona and the participants of the user study from the University of Arizona Medical Center.

This project was sponsored by the following grant: NIH/NLM, "UMLS Enhanced Dynamic Agents to Manage Medical Knowledge", 1 R01 LM06319-01Al, February 2001–February 2004.

REFERENCES

[1] O. Baujard, V. Baujard, S. Aurel, C. Boyer, R.D. Appel, Trends in medical information retrieval on internet, Comput. Biol. Med. (1998) 589–601.
[2] L. Bin, K.C. Lun, The retrieval effectiveness of medical information on the web, Int. J. Med. Inform. 62 (2001) 155–163.
[3] L.T. Su, A comprehensive and systematic model of user evaluation of web search engines: I. Theory and background, J. Am. Soc. Inform. Sci. Technol. 54 (2003) 1175–1192.
[4] A. Gaudinat, P. Ruch, M. Joubert, P. Uziel, A. Strauss, M. Thonnet, R. Baud, S. Spahni, P. Weber, J. Bonal, C. Boyer, M. Fieschi, A. Geissbuhler, Health search engine with e-document analysis for reliable search results, Int. J. Med. Inform. 75 (2006) 73–85.

[5] A. Sutcliffe, M. Ennis, Towards a cognitive theory of information retrieval, Interact. Comput. 10 (1998) 321–351.
[6] C. Kuhlthau, Longitudinal case studies of the information search process of users in libraries, Library Inform. Sci. Res. 10 (1988) 257–304.
[7] G. Marchionini, Information Seeking in Electronic Environments, Cambridge University Press, 1995.
[8] E.F. de Lima, J.O. Pedersen, Phrase recognition and expansion for short, precision-biased queries based on a query log, in: Presented at the Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, 1999.
[9] N.C.M. Ross, D. Wolfram, End user searching on the internet: an analysis of term pair topics submitted to the excite search engine, J. Am. Soc. Inform. Sci. 51 (2000) 949–958.
[10] A. Spink, D. Wolfram, M.B.J. Jansen, T. Saracevic, Searching the web: the public and their queries, J. Am. Soc. Inform. Sci. Technol. 52 (2001) 226–234.
[11] E.G. Toms, R.W. Kopak, J. Bartlett, L. Freund, Selecting versus describing: a preliminary analysis of the efficacy of categories in exploring the web, in: Presented at the Proceedings of the 10th Text REtrieval Conference (TREC 2001), Maryland, 2001.
[12] T. Lau, E. Horvitz, Patterns of search: analyzing and modeling web query refinement, in: Presented at the Proceedings of the Seventh International Conference on User Modeling, 1998.
[13] D. Hawking, N. Craswell, Overview of the TREC-2001 web track (TREC 2001), in: Presented at the Proceedings of the 10th Text REtrieval Conference, 2001.
[14] W.R. Hersh, S. Price, L. Donohoe, Assessing thesaurus-based query expansion using the UMLS metathesaurus, in: Presented at the Proceedings of the 2000 Annual AMIA Fall Symposium, 2000.
[15] J.C. French, A.L. Powell, F.G.N. Perelman, Exploiting a controlled vocabulary to improve collection selection and retrieval effectiveness, in: Presented at the Proceedings of the 10th International Conference on Information and Knowledge Management, Atlanta, Georgia, 2001.
[16] J. Koenemann, N.J. Belkin, A case for interaction: a study of interactive information retrieval behavior and effectiveness, in: Presented at the Proceedings of the Conference on Human Factors in Computing Systems, Vancouver, Canada, 1996.
[17] H. Joho, C. Coverson, M. Sanderson, M. Beaulieu, Hierarchical presentation of expansion terms, in: Presented at the Proceedings of the ACM Symposium on Applied Computing, Madrid, Spain, 2002.
[18] M.L. Resnick, M.W. Vaughan, Best practices and future visions for search user interfaces, J. Am. Soc. Inform. Sci. Technol. (2006), vol. (Early View).
[19] Y. Nemeth, B. Shapira, M. Taeib-Maimon, Evaluation of the real and perceived value of automatic and interactive query expansion, in: Presented at the Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval (SIGIR'04), Sheffield, South Yorkshire, UK, 2004.
[20] B. Jansen, A. Spink, T. Saracevic, Real Life, real users, and real needs: a study and analysis of user queries on the web, Inform. Process. Manage. 36 (2000) 207–227.
[21] A.T. McCray, T. Tse, Understanding search failures in consumer health information systems, in: Presented at the Proceedings of the AMIA 2003 Symposium, Washington, DC, 2003.
[22] I. Ruthven, Re-examining the potential effectiveness of interactive query expansion, in: Presented at the Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'03), Toronto, Canada, 2003.
[23] S. Greene, G. Marchionini, C. Plaisant, B. Shneiderman, Previews and overviews in digital libraries: designing surrogates to support visual information seeking, J. Am. Soc. Inform. Sci. 51 (2000) 380–393.
[24] G. Marchionini, C. Plaisant, A. Komlodi, Interfaces and tools for the library of congress national digital library program, Inform. Process. Manage. 34 (1998) 535–555.
[25] E. Amitay, C. Paris, Automatically summarising web sites: is there a way around it? in: Presented at the Proceedings of the Ninth International Conference on Information and Knowledge Management, McLean, Virginia, 2000.
[26] A. Woodruff, R. Rosenholtz, J.B. Morrison, A. Faulring, P. Pirolli, A comparison of the use of text summaries, plain thumbnails, and enhanced thumbnails for web search tasks J. Am. Soc. Inform. Sci. Technol. 53 (2002) 172–185.
[27] B. Shneiderman, The eyes have it: a task by data type taxonomy for information visualizations, in: Presented at the Proceedings of the IEEE Symposium on Visual Languages, 1996.
[28] W. Pratt, L. Fagan, The usefulness of dynamically categorizing search results, J. Am. Med. Inform. Assoc. 7 (2000) 605–617.
[29] R. Grishman, S. Huttunen, R. Yangarber, Information extraction for enhanced access to disease outbreak reports, J. Biomed. Inform. 35 (2002) 236–246.
[30] G. Leroy, H. Chen, Genescene: an ontology-enhanced integration of linguistic and co-occurrence based relations in biomedical texts, J. Am. Soc. Inform. Sci. Technol. (Special Issue) 56 (2005) 457–468.
[31] J. Beier, T. Tesche, Navigation and interaction in medical knowledge spaces using topic maps, Int. Congr. Ser. 1230 (2001) 384–388.
[32] G. Leroy, A.M. Lally, H. Chen, The use of dynamic contexts to improve casual internet searching, ACM Trans. Inform. Sys. 21 (2003) 229–253.
[33] E.M. Thury, Analysis of student web browsing behavior: implications for designing and evaluating web sites, in: Presented at the Proceedings of the 16th Annual International Conference on Computer Documentation, Quebec, Canada, 1998.
[34] H. Thimbleby, Permissive user interfaces, Int. J. Hum. Comput. Stud. 54 (2001) 333–350.
[35] G. Leroy, H. Chen, Meeting medical terminology needs: the ontology-enhanced medical concept mapper, IEEE Trans. Inform. Technol. Biomed. 5 (2001) 261–270.
[36] G. Leroy, H. Chen, MedTextus: an ontology-enhanced medical portal, in: Presented at the Proceedings of the Workshop on Information Technology and Systems (WITS), Barcelona, 2002.
[37] A.T. McCray, Informatics research, development, and training at the Lister Hill national center for biomedical communications, in: R. Haux, C. Kulikowski (Eds.), Yearbook of Medical Informatics 2003, IMIA, 2003, pp. 193–203.
[38] J.R. Lewis, IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use, Int. J. Hum. Comput. Interact. 7 (1995) 57–78.
[39] B.J. Fogg, J. Marshall, A. Osipovich, C. Verma, O. Laraki, N. Fang, et al., Elements that affect web credibility: early results from a self-report study, Paper Presented at CHI 1–6 April 2000.
[40] B.J. Fogg, C. Soohoo, D.R. Danielson, L. Marable, J. Stanford, E.R. Tauber, How do users evaluate the credibility of web sites? A study with over 2500 participants, Paper Presented at the 2003 Conference on Designing for User Experiences, San Francisco, California, 2003.
[41] T. Kohonen, Self-organizing maps, Springer-Verlag, Berlin, 1995.