Jason K. Sokoloff. Search Patterns Through a Health-Information Site: Considering the Need for Complex Subject Indexing. A Master's Paper for the M.S. in L.S degree. April, 2006. 30 pages. Advisor: Jane Greenberg

This study considers the impact of taxonomy development on user query-expansion patterns at NC Health Info, a Web database of North Carolina online health and medical resources. In consideration of simplifying NC Health Info's taxonomy, user session logs were analyzed for selection frequency of general and specific topics and directional patterns between general and specific topics as initial and subsequent selectors. Based on a sampling of session logs over a seven-month period, users exhibited no clear preference for general or specific topics. In an analysis of topics deemed crucial to North Carolinians by a governor's task force, patterns illustrated a significant preference for specific topics over general topics. This research, and the results of previous studies regarding taxonomy development and query-expansion, suggests that a simple taxonomy would less effectively serve users.

Headings:

Consumer health information

Thesauri

Query expansion

SEARCH PATTERNS THROUGH A HEALTH-INFORMATION SITE: CONSIDERING THE NEED FOR COMPLEX SUBJECT INDEXING

by
Jason K. Sokoloff

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Library Science.

Chapel Hill, North Carolina

April 2006

Approved by

_____

Jane Greenberg

TABLE OF CONTENTS

INTRODUCTION

In a traditional library setting, information seekers have the benefit of interacting with information professionals who are trained to converse, probe and guide patrons through a well-known process known as the reference interview. By answering direct questions, observing physical cues and even noticing questions not being asked, the reference librarian is effectively able to guide information seekers to the sources they need. But as information seekers increasingly turn to online sources for research, a challenge emerges. How to guide users through rich online information sources has become an important topic in the field of information-retrieval research. Particularly in situations where information seekers wish to be discreet, the question of how to best guide them may be likened to the challenge of aiding library patrons who bypass the reference desk and head straight into the stacks. Guiding information seekers in these situations are term vocabularies. By examining query-expansion patterns in one particular health-information Web directory, NC Health Info, this research will consider the need for simple and complex subject taxonomies.

According to the Pew Internet & American Life Project, some 79 percent of all U.S. adult Internet users — or 95 million Americans — have gone online to search for health and medical information (Fox, 2005). Averaging some 7 million adult users day, health-information seekers are greater in number that those who purchase products through Internet retailers (Rainie & Horrigan, 2005). The Pew investigators and other have explained that as Internet users become more experienced and as high-speed access

becomes more prevalent, information seekers are becoming increasingly reliant on and expectant of quality information on the Web. These increasing levels of comfort and accessibility coincide with a growing number of thorough and reliable medical information sources online, from government-sponsored public-education initiatives to pharmaceutical consumer-information pages. (Fox & Rainie, 2000)

Ensuring the productivity of consumer health-information seekers and the success of their research efforts involves the study of two particular aspects of information retrieval: (1) how consumers search for health information, and (2) how indexing systems provide guidance in topical research.

In the last twenty-five years, observations regarding consumer health-information seeking were primarily addressed in one of two ways: (1) marginally, in relation to discussions of librarian responsibilities, ethics and legal issues, as in (Eakin, Jackson, & Hannigan, 1980); and (2) narrowly, in consideration of very specific library-patron populations like African American women (Gollop, 1997) or the elderly (Campbell & Nolfi, 2005). Since the mid-1990s and the growth of consumers as Internet users, researchers had the benefit of observing information-seeking patterns through the navigation of electronic and online information resources.

Related to more recent research in information-seeking behavior is the use of controlled vocabularies. Particularly among medical and scientific information sources, controlled vocabularies are staple tools of indexers and researchers that aid experienced researchers in locating precise information. In an effort to ensure consistent indexing and successful searching, taxonomies like the National Library of Medicine's Medical Subject Headings (MeSH) and the biomedical/pharmacological EMTREE are comprised

of a topical hierarchy of terms referenced by a thesaurus of related terms. Now that nonprofessionals are offered similar levels of information access as skilled researchers, a challenge has emerged: how to make these highly developed vocabularies work for less sophisticated information seekers.

The research presented in this paper will examine the topic-selection patterns by health-information seekers in the context of a particular health-information database: a Web site called NC Health Info (www.nchealthinfo.org). A joint project of the Health Sciences Library and School of Information and Library Science at the University of North Carolina at Chapel Hill, NC Health Info (NCHI) is an index of local health resources on the Web. NCHI uses a simple interface that allows users to easily locate health and medical Web pages of local geographic relevance.

In an effort to make NCHI's indexing processes more efficient, consideration has been given to the notion of simplifying NCHI's taxonomy. Driven by this consideration, the following research questions will address four related user behaviors:

1. Is there a clear user preference for general or specific topics?
2. Is there a clear user preference for query-expansion patterns in the direction of general-to-specific of specific-to-broad?
3. Is there a preference for general or specific topics among particular types of topics?
4. Is there a relationship between topic-selection patterns overall and topic-selection among particular topic types?

Previous research related to this analysis generally falls into three realms: consumer health-information searching preferences and behavior, effectiveness of

medical taxonomies, and the query-expansion process — all of which seem to suggest that complex taxonomies dependent on rich thesauri, comprised of specific terminology, are more favorable in promoting information retrieval. NCHI's structure and processes are unique from those considered in existing studies — enough so that explicit consideration of its distinct user patterns is necessary to make firm conclusions.

LITERATURE REVIEW

Three areas of study may be considered in relation to the issues facing NC Health Info. First, the observed habits and preferences of consumers searching for health information helps to understand the types of topics that users may be likely to select. Secondly, a consideration of previous work on medical taxonomy effectiveness is useful for comparative purposes. Lastly, a look at vocabulary thesauri and their impact on query expansion will help to understand topic-selection patterns.

Consumers Seeking Health Information

While the NCHI format is somewhat unique among consumer health-information resources, the results of several well-documented studies provide some guidance for NCHI in consideration of altering its taxonomy.

The Pew Internet & American Life Project, an organization that regularly conducts research concerning how Internet usage and implications, conducted in 2000 a study that remains a definitive source for the nuances of consumer health-information seeking behavior. Related to the concept of narrow or broad topic selection, the

researchers were clear: consumers are by far more interested in researching on specific topics. Of all health-information seekers surveyed, 70% said they went online to locate information about a particular health disorder or disease. In support of this measure, the survey illustrated two common scenarios of health-information seekers: (1) a patient who has been recently diagnosed, and (2) a friend or family member of a recently diagnosed patient. Given this common catalyst for health-information searching, the authors reported that "The health seekers in our survey were more likely to be focused on an immediate problem, rather than general information…" (Fox & Rainie, 2000).

A similar report by (Poensgen & Larson, 2001) for management advisory firm Boston Consulting Group, involved the study of European consumers in a series of focus groups. Though the objectives of their efforts were commercially driven — Boston Consulting was appealing to potential health-care information industry clients — the findings regarding topic scope are in concert with other, more academic analyses have revealed. Poensgen and Larson wrote: "Patients want data and deep content on a specific disease or condition…. [They] want very focused and detailed information about their specific condition or disease." (p. 13).

These views of what consumers seek in health-information research seem incompatible with the idea of flattening NC Health Info's taxonomy. While revealing, these behavioral studies don't make particular mention of how users go about searching these narrow topics. Where these studies become less relevant, an ample base of existing research in the realm of medical taxonomy usage, can provide further direction to the NCHI considerations.

Medical Taxonomies

Researchers examining the use and effectiveness of medical vocabularies typically consider such use by health and information professionals. But even in cases of such taxonomies being engaged by consumers, the consensus recommendation appears to be the favor complex structures.

Zeng, Kogan, Ash, Greenes, & Boxwala, (2002) suggested that consumers generally have a difficult time in successfully searching health information because of their inherent lack of comprehension of medical terminology. Their study involved the analysis of Web usage logs and follow-up interviews with consumers accessing two health-information Web sites. The findings indicated that consumers were generally not very accomplished at retrieving what they needed. Zeng et al. (2002) recommended that a comprehensive terminology — one that connects professional and lay terms — is a necessity for any system aiming to serve consumers.

Experimenting with bridging this gap between professional and consumer terminology, (Patrick, Monga, Sievert, Houston Hall, & Longo, 2001) experimented with the a developed controlled vocabulary that combined a medical thesaurus with a dictionary of regional American English. Using terms related to the subject of diabetes, the authors found the complex combination yielded significant results in terms of matching consumer-oriented and professional-oriented references.

Even discussion attributing the benefits of an abbreviated structure stresses the need for complex development. In a recount of how the American Medical Association's publishing division developed a concise taxonomy, McGregor described his challenge of creating a structure that would cover topics covered in major medical journals yet also fit

on a single Web page. McGregor and his colleagues implemented a facets approach to this task, and while he reported the end-result a success, he suggested that the system's qualities are still dependent on rich subject-heading terminology already embedded in the markup language of the journal articles being indexed by the more concise system (McGregor, 2005)

Thesauri and Query Expansion

At work in NC Health Info is a controlled vocabulary that, while itself elaborate, is relatively simple in comparison to the MeSH taxonomy on which it is based. Essentially comprised of two levels of general terms and specific terms (along with a significant set of terms that are neither general or specific), NCHI's thesaurus directs from layman terms to clinical terms, as is expectant of any effective controlled vocabulary system. An additional feature, though, directs users to general topics after they have selected specific terms. (This direction appears in the form of a dynamically generated sidebar link labeled "Related Topics.") This functionality enables users in the act of "query expansion," another research subject that has received significant attention among information-retrieval researchers and has relevance to NCHI's considerations.

In his seminal review on query expansion, Efthimiadis (1996) defined the process as one where an initial query is modified with subsequent terms. This modification can occur in the form of added or removed terminology and may result from automatic, manual or interactive functions. That is, query expansion can be based wholly on system-generated processes where topic selection is constrained to predetermined choices or involve some level of user contribution, as with a traditional, free-text Boolean searching.

Efthimiadis' work, a thorough examination of the effectiveness of different types of thesauri serving query expansion, provides some clear distinctions and comparisons among previously researched controlled vocabularies.

Building upon this previous research outlined by Efthimiadis, other researchers have since explored the ideal implementations of thesauri and controlled vocabularies as they serve varying query-expansion methods. Greenberg (2001) found that narrow-term and synonym references were better suited for automatic query expansion while related terms were better suited for interactive query expansion. Greenberg speculated that broad terms were also better candidates for interactive query expansion, though the experiment at hand yielded insufficient data to make a firm determination.

In an examination of query expansion and user behavior in an agricultural thesaurus, Shiri & Revie (2006) reported that more complex topics necessarily yielded more query expansion and that the process was enlightening for about half of the subjects tested, where users reported having learned new terms as a result of query expansion.

Query expansion research, then, seems to indicate that query expansion is desirable and useful, particularly in areas of complex terminology. Again, the previous work in this area doesn't support the case for simplifying NCHI's taxonomy. But again, NCHI thesaurus and query-expansion patterns don't exactly fit into the established models.

Based on Efthimiadis' classifications of query-expansion methods, it is apparent that NC Health Info falls somewhere between the categories of "interactive" and "automatic." Neither is a completely accurate label, since users are only prompted to expand their query when they have initially selected a specific health term. If a user

selects “lung cancer,” for example, she is presented in a sidebar link the suggested “related topic” (in actuality, a more general topic) of “Cancer.” This pattern does not work in a reverse order, presented narrower “related” topics when a user has initially selected a general one.

NCHI is also anomalous among previous studies in its simple taxonomy. Fundamentally based on MeSH headings, NC Health Info’s thesaurus presents users with the option of selecting from health topics that are either general or specific. (There are also a significant number of terms that are neither general nor specific.) So-called related topics, typically where a layman term is followed by a “see” reference to more precise clinical terminology isn’t exactly automatic. Users are not constrained by any automated process to follow the “see ” reference, but there is no alternative except to select an altogether different topic.

NC HEALTH INFO OVERVIEW

The User Session

NC Health Info enables users to locate local health resources in a simple, two-step process: (1) select a topic from a predefined list of health and medical terms, then (2) select from a list of North Carolina cities or counties. Upon receipt of these items, the NCHI database returns the user to a dynamically generated list of on-topic and geographically relevant Web resources. (See Figure 1.)

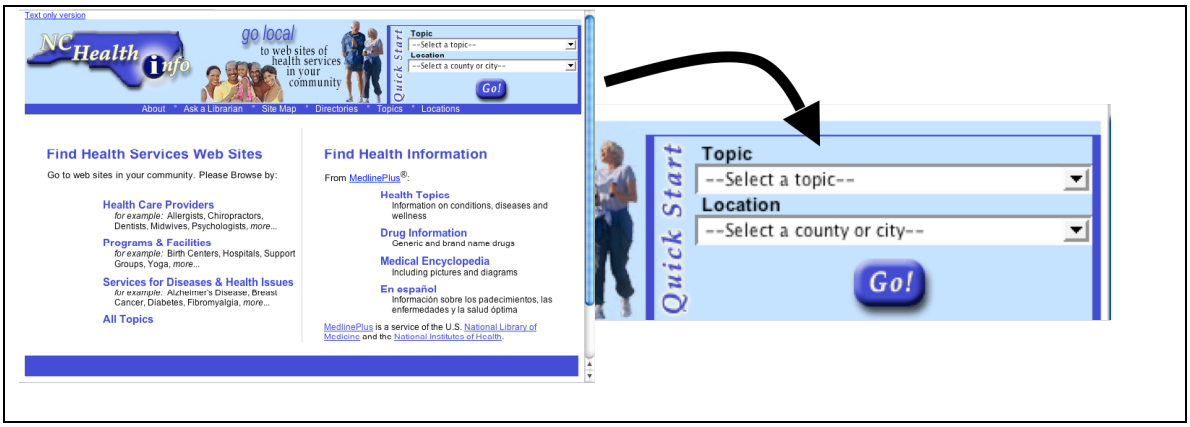

Figure 1: The NC Health Info home page (left) and a close-up of the quick-start topic and location selection menu.

An example: A user logs on to NCHI and selects "Children's Health" from the topic menu and "Watauga County" from the location menu. NCHI then delivers a page of resource listings, arranged into topic groupings, like family practitioners, pediatricians, children's hospitals, audiologists, mental health professionals, and other specialists or organizations that provide medical services to children in Watauga County. Each listing is comprised of the resource name linked to the provider's URL, the parent organization (if applicable), and a contact phone number.

Worth mentioning, although not central to the research at hand, is the fact that in a sidebar alongside these linked entries appear another set of related links to health information on MedlinePlus (www.medlineplus.gov) Web site. Once the example user has found her local resources, she may now follow these links to learn more about particular aspects of children's health from the volumes of information made freely available by the National Library of Medicine (NLM) and the National Institutes of Health. (See Figure 2.)

Figure 2: A sample NCHI results page, illustrating links for Children's Health resources in Watauga County. The right column contains links to MedlinePlus.

This feature is an added convenience for NCHI users but is also an indicator of an important relationship between NCHI and NLM, which funded through a grant the database's early development. The added-value links are one side of a reciprocal relationship. When browsing MedlinePlus information, users are presented with a corresponding link, labeled "Go Local," that links to state-specific providers relevant to the topic of the current page. Users may currently select from about a dozen states, but North Carolina was the first available as it has served as the pilot project in similar indexes now operating around the country.

Maintaining Database Records

The NC Health Info database currently contains more than 3,000 records, providing users access to Web pages of health and medical providers across the state's 100 counties. While impressive, the project has room to grow. Resources serving the state's population centers are well developed, but less improvement are resources serving rural areas and resources that may have no Web sites at all. The tasks of cultivating such resources have received low priority, though. The key obstacle is the amount of work currently necessary for merely maintaining the accuracy of the current database.

Since its conceptual stages, NCHI has always been a librarian-lead project. Rather than a mere accumulation of health care Web sites, NCHI was developed with the same discretion and scrutiny that might be applied in traditional library collection development. This approach makes what otherwise might be just a Web directory into an authoritative, meticulously organized and reliable "collection of links," as Health Sciences Library directors prefer to label it.

Central to this level of authority is the current process of human indexing Web sites. The work involved is substantial, requiring the attention and decision-making of a project director (a full-time professional librarian) assisted by two library-science graduate students who serve as catalogers. Using a Web-form interface custom designed for NCHI, catalogers create detailed electronic records for each resource. Each record contains the information that displays to the end user: resource name, city and county location, telephone number and URL. Invisible to users is the elaborate subject-metadata indexing system, where catalogers apply health topics from a hierarchical tree of local terms and health-topic pairings patterned after the National Library of Medicine's Medical Subject Headings taxonomy. The end-result then is a user interface that appears

to enable faceted indexing, even though a complex hierarchy is operating behind the scenes.

To ensure continued currency and accuracy, catalogers review each record at least annually, and it is this maintenance process that poses an obstacle for NC Health Info's further development. Reviewing a resource entails the comparison of the previously cataloged record to the current Web page and the subsequent incorporation of any necessary revisions to the record metadata. With the sheer number of records to review, this resource verification process has eclipsed other, more creative efforts that would increase the service's scope and usefulness. As is the case with many free, non-revenue generating, publicly funded services, NCHI's budget is modest and contains no certain, long-term allowances for staff expansion (or even maintaining current staffing levels).

While professional-level record maintenance has been central to the concept and theme of NCHI, the amount of work necessary for its upkeep requires consideration of more efficient methods. One such modification involves the flattening of the existing subject taxonomy, one that currently requires catalogers to select from nearly 8,000 possible pairings of health topics with more general health subject terms. A compressed taxonomy would conceivably streamline the indexing process, but consideration must first be made on how a simplified structure would impact NCHI users.

By examining patterns in existing user topic selection, this research will consider the potential impact of altering the current taxonomy. Particular attention will be paid to user tendencies toward general or specific topics, and the order in which they move between general and specific topics. ("General" and "specific" will be used interchangeably in this discussion with the descriptions "broad" and "narrow" in relation

to NCHI thesaurus terminology.) In exemplary terms, this research will consider whether a user is more likely to select "Cancer" or "Lung Cancer," a type of cancer, and in what order he is likely to select these terms as he moves through his NCHI session. Though these patterns may be somewhat dictated by parameters of the system, topic selection is primarily user-controlled.

PROBLEM STATEMENTS

Flattening the existing NCHI taxonomy essentially means removing specific topics and relying solely on more general ones. (Specific and general topics will hereafter also be described as "narrow" and "broad.") The decision to restructure in this manner requires consideration of current user topic-selection patterns and the apparent popularity of particular topics. From this rationale emerge two related research questions:

1. Among NC Health Info users, is there a clear preference for broad or narrow topics selected?
2. Is there a clearly preferred query-expansion pattern in terms of direction of broad-to-narrow or narrow-to-broad topics?

The results, as described below, triggered two subsequent questions that would aid in a final recommendation for taxonomy development:

3. Is there a particular preference for narrow or broad topics among particular types of topics?

4. Is there a relationship between topic-selection patterns and type of topic selected? (As detailed in the following section, topic type in this case refers to the deemed importance of particular topics over others.)

RESEARCH METHOD

A log analysis was employed to answer the research questions and establish a rationale for altering the NCHI taxonomy (or to make a case for retaining it in its current state). NCHI planners make use of ongoing session logs that track a variety of user movements through the site, including page view statistics, entry and exit points, and topic search patterns. This session-log tracking is manifested in a more than two-dozen reports, groupings of statistics to reflect particular aspects of usage patterns. For the purpose of this project, two such reports were run. Both entailed the tallies of user-selected topics. The resulting topics were considered for their "broadness" (general) or their "narrowness" (specific), a key indicator related to the complexity of NCHI's taxonomy.

The first utilized report, titled "Resource Searches by Health Topic," illustrates the number of times that individual NCHI health topics are selected. This report was run for a period of seven months, from January 1 to July 31 of 2005, and sorted to reveal the most-selected topics for the entire seven-month time period. Each of the top 100 topics were then labeled as broad or narrow, a decision-making process based primarily on knowledge of the health topic classification structures in general and, secondarily, knowledge of the NCHI cataloging taxonomy in particular.

The second report, "Resource Search Patterns" illustrates trends in pairings of user topic selections in terms of "first topic selected" and subsequent topic selected." In the vast majority of cases, no clear connection could be made between the two adjacent topics, so the report was run only for randomly selected days for each of the seven months previously considered. A random-number generation table was used to determine which days to examine, and only pairings that occurred more than once were considered. Where a connection could be made between the adjacent topics, a determination was made as to the direction of the user movement: broad-to-narrow or narrow-to-broad. Again, this judgment did not entail any automation but was instead based on researcher knowledge of health-topic subjects and their hierarchical arrangements.

Research questions 1 and 2 required only a tally of broad versus narrow (in set one) and broad-to-narrow versus narrow-to-broad (in set two). Given the results of testing these initial questions, it became evident that a look at topic-selection trends would be beneficial. Rather than considering patterns among all topics or even the most-selected topics, the next question to emerge dealt with health topics of a particular quality. To consider question 3, the same daily "pattern" logs were filtered for pairings that pertained to one of six chronic-disease categories: (1) arthritis and osteoporosis, (2) asthma, (3) cancer, (4) diabetes, (5) heart and circulatory disease, and (6) obesity. These areas were selected for their importance to North Carolinians, as deemed by a 1999 gubernatorial task force on ten-year health objectives (Bobbitt-Cooke, 1999). These "Healthy Carolinian" topics, so named for the task force that identified them, have served NCHI in its development and maintenance. Catalogers heed resources that treat these medical conditions, giving them priority and extra scrutiny.

The last step entailed the question of whether there was a significant relationship between the two sets of topic-selection patterns, the Healthy Carolinian topics and "all topics." From the one-day-per-month sampling of query-expansion patterns, pairings involving one or more Healthy Carolinian topics were identified for this comparison. To consider question 4, a chi-square test was run on these frequencies to determine a connection between them.

DATA ANALYSIS

A series of four session-log data sets were examined for evidence of trends in broad and narrow topic selection. The initial findings revealed no significant preference for narrow or broad topics and no clear inclination in query-expansion patterns in terms of broad-to-narrow or narrow-to-broad. Subsequent results indicated a significant preference for narrow topics among "crucial" health topics — the so-called "Healthy Carolinian" topics deemed critical for the state of North Carolina by a gubernatorial task force.

Hypothesis 1: There is a significant difference between the number of times that broad topics and narrow topics are chosen.

In order to consider the overall popularity of broad topic selection versus narrow topic selection, NCHI web logs were parsed for the top 100 selected topics in a seven-month period. Once these top 100 were collected, each topic was assigned a label of broad or narrow, depending on the scope of the topic and its place in the NCHI and

MeSH taxonomic structures. The end result, as illustrated in Table 1, is a null hypothesis. The difference between narrow and broad topics is negligible, with 39% of the most-selected topics being broad and 42% being narrow.

Table 1. Comparison of most-selected NCHI health topics, Jan. to July, 2005

Topic Scope	No. of Times Selected
Broad	39
Narrow	42
Neither	19

Sample: top 100 selected topics

Hypothesis 2: There are discernable preferences in topic query-expansion patterns of in terms of movement from broad-to-narrow or narrow-to-broad.

This sampling entailed running a daily pattern report for a randomly selected day for each month starting in January and ending in July of 2005. The pattern report revealed every instance of a pairing of first and subsequent topic selected. In order to ensure a manageable number of pairings and to reflect the likelihood of more relevant patterns, each daily list was edited to include only those pairings that occurred two or more times each day. For these pairings, a label of broad-to-narrow or narrow-to-broad was applied, based again on the researcher's knowledge of hierarchical relationships between topics.

This test also required only a tally of the results. As shown in Table 2, when totaled, the number of opposite directional movements was exactly the same. That is, for the randomly selected days, the number of times that users moved in a broad-to-narrow

direction was the same as the number of narrow-to-broad movements: 38 times each, with 201 two-step movements with no apparent relationship. Another hypothesis disproved.

Table 2. Comparison of NCHI topic-selection patterns, Jan. to July, 2005

Pattern	No. of Times Selected
narrow → broad	38
broad → narrow	38
no clear relationship	201

Sample: one day per month

With two null hypotheses, an opportunity emerged to examine topics in a slightly different manner. It was becoming clear that omitting narrow topics from the taxonomy could be counterproductive for users. A subsequent hypothesis considers the importance of narrow topics among more "important" topics — that is, the Healthy Carolinian chronic-disease topics deemed crucial to North Carolina by the aforementioned governor's task force. With the difference between narrow and broad topic-selection patterns so far appearing to be marginal, this further consideration of patterns among this special set of topics might reveal more vividly the significance of narrow topics for particular subject areas.

Hypothesis 3: Among crucial "Healthy Carolinian" medical subjects, there is a significant difference between the number of times that broad topics and narrow topics are chosen.

The tally of directional pairings among the Healthy Carolinian topics illustrated a clear preference for the movement of narrow-to-broad in first and subsequent topic selections. As displayed in Table 3, users moved in this direction 48 times and only 10 times in a broad-to-narrow pattern.

Table 3. Comparison of NCHI "Healthy Carolinian" topic-selection patterns, Jan. to July, 2005

Pattern	No. of Times Selected
narrow → broad	48
broad → narrow	10

Sample: one day per month

Hypothesis 4: There is a relationship between topic-selection patterns and type of topic (all topics versus Healthy Carolinian topics).

In order to determine a connection between the patterns observed among Healthy Carolinian topics and all NCHI topics, a chi-square test was performed to compare the calculated frequencies. With a region of rejection greater than or equal to 3.84, the test demonstrated that there was a statistically probable relationship between the two sets of frequencies and that the variation was not due merely to chance. Table 4 contains each of the data sets and the parameters and results of the chi-square test.

Table 4. Comparison of topic-selection patterns across topic types, Jan. to July, 2005

Pattern	No. of Times Selected: Healthy Carolinian Topics	No. of Times Selected: All Topics
narrow → broad	48	38
broad → narrow	10	38

Results of chi-square test: $\chi 2 = 4.95$ on 1 *df;* $p = 0.5$.
Sample: one day per month

DISCUSSION

Topic-Selection Pattern Assumptions

It is important to note the inherent assumptions that factor into the interpretation of results and subsequent recommendations discussed in sections that follow. The "Resource Search Patterns" report illustrates pairings of topics and subsequent topics selected by users in the span of a single session. The interpretation of a subsequent selection is that it marks the continuation of a search. When a search pattern moves from a broad topic to a narrow topic, the inference is that the user required a more specific label to hone in on the subject of interest. With narrow-to-broad patterns, the presumption is that users are better served by the more general label.

Following are explicit examples of how these patterns might manifest in a typical user session. (The estimated tallies indicated were accurate as of March 12, 2006., the last time these examples were run.)

- *Narrow-to-broad*: (1) User selects broad topic "Cancer" in Mecklenburg County and receives some 80 resulting links classified into a number of

subtopics like cancer clinics, oncologists, hospitals, pain clinics, radiologists and surgeons. (2) Faced with the prospect of too many listings to be convenient, the user selects the more narrow "Lung Cancer" in an effort to hone in on very specific resources. The new topic for the same geographic region has been whittled down to 12 resulting resources.

- *Broad-to-narrow:* (1) User selects narrow topic "Irritable Bowel Syndrome" in Mecklenburg County and receives one result: an acupuncture clinic. (2) User widens the topic to the more general "Digestive Diseases" to generate a list of 18 relevant links, most of them for gastroenterologists.

User-Driven and System-Driven Topic Selection

It is also important to note that the move from narrow to broad is not always entirely based on the user's own needs. Rather, the functionality of NCHI guides users towards particular topics under certain circumstances. One example of this occurs when a user selects a narrow topic that, according to the behind-the-scenes taxonomy, has no resources associated for the selected county. The result is a generated list of statewide or neighboring-county resources in lieu of a "no results found" message. Another possibility involves an NCHI convention where a "Related Topics" link appears in a sidebar after a narrow topic has been selected. Both of these situations illustrate where NCHI lacks information and drive user selection. If a user first chooses "Epilepsy," for instance, he is offered alongside his results a link to the broader "Neurological Disorders." Had the list

of results for “Epilepsy” been insufficient, the user is prompted to subsequently select a broader term.

Had the statistical comparison of broad and narrow terms and patterns been marginally distinct, there would be cause to more closely examine the approximate level of influence that the system influences user topic selection. As it stands, the statistics illustrate a considerable difference between topics of varying scope and the movement between them.

The data concerning topic-selection frequency is clear: NCHI users are relying substantially on narrow health topics in their search for resources. The Web metrics analysis found that ratio of broad to narrow topic selections was generally one-to-one. An examination of the trends in first and subsequent topic selection patterns revealed that, again, the movement in either direction was about equal.

Had either of these data sets proven to be more one-sided, showing a significant disparity, a stronger case could have been made for the removal or retention of narrow topics. As it stands, the use of narrow and broad topics is about equal across all topics.

By examining the pattern frequencies in a more qualitative way — considering topics in terms of type — a new trend emerged, one that indicated a strong preference for narrow topics among more crucial topic categories. The chi-square test supported the probability that this variance was statistically sound and not a chance occurrence..

CONCLUSION & RECOMMENDATIONS

This research addressed the possible impact of a simplified subject taxonomy on guiding NC Health Info users. In summary, the session-log analysis first revealed two user topic-selection trends: (1) no clear preference for broad or narrow topics; (2) no clear preference for query-expansion patterns in terms of broad-to-narrow or narrow-to-broad. Further examination illustrated a preference for specific topics among Healthy Carolinian topics deemed crucial to the state of North Carolina. Finally, a chi-square calculation revealed a statistically significant relationship between query-expansion patterns and types of topics — that is, Healthy Carolinian versus all other topics.

These findings, combined with the conclusions and recommendations of previous work outlined in the literature review, suggest that a compressed taxonomy would be a disservice to NCHI users. The following discussion considers possible scenarios for further consideration in NCHI's development.

Simplifying the Entire Taxonomy

At the center of this research was the basic question of whether a case could be made for flattening the NCHI taxonomy in an effort to make the indexing process more efficient. An analysis of topic selection patterns suggests that the answer is no. The current use of narrow topics is significant and to remove them from the indexing structure would be doing a disservice to NCHI users.

When NCHI users select from the more crucial "Healthy Carolinian" topics, the preference for narrow topics is even more pronounced, indicating that certain topics require the inclusion of narrow-term metadata in order to be properly located by users.

Simplifying Segments of the Taxonomy

Perhaps an inverse look at the crucial-topic trend could serve NCHI's needs. The rationale is that if critical health topics (i.e. Healthy Carolinian topics) most require complex taxonomies, perhaps less crucial topics would suffice with flatter structures.

Examining this possibility would require further analysis of search pattern trends, but not before determining the relative importance of topics overall. The conducted research benefited from the Healthy Carolinians task force's work on identifying crucial topics, but a similar selection of lesser topics would require some difficult choices. A process involving the relegation of certain topics to the bottom of a list would be needed — not an impossible task, but one that would require some level of objectivity. It might be easy to immediately label non-medical topics as less important. The line is easily drawn between diabetes and yoga, for instance. But the job becomes more difficult in ranking some of the social well-being topics that populate NCHI's taxonomy. At the very least, a test could be performed on a small percentage of obviously less crucial topics — rare diseases and issues marginally related to health and medicine, for instance. If search patterns appear to be more broad leaning, as a new hypothesis would pose, a case could be made for flattening the taxonomy for these less crucial categories while more important topic areas would remain in their current state.

Alternatives for Consideration

A number of other possibilities exist for the future development of NC Health Info, many of which would require significant time and effort to pursue but could

potentially aid in overcoming the current burdens associated with record verification. One idea of which NCHI directors are already aware involves an automated process of verifying record accuracy. Such an application would necessitate original development or at least the tailoring of an existing system but could potentially save the time human catalogers spend comparing Web pages to record metadata.

Another solution, one that NCHI is currently pursuing, is the involvement of health care providers in the maintenance of their own records. The task of record verification, although time consuming, is a highly effective, where catalogers automatically receive notification when a resource is due for its annual review. Building upon this established procedure, NCHI staffers are currently experimenting with a tweak of the system that would deliver a notification via email to the health care provider that requests their participation in ensuring the record's accuracy.

Other ideas might have less immediate impact on the cataloging workload but could prove useful to users and to NCHI planners in their understanding of how users are engaging the service. The idea of implementing a search engine is one that has been discussed at NCHI but has never evolved beyond the conceptual stage.

A related possibility would be the incorporation of a visible taxonomy. This concept would make available the same structure visible to catalogers, visible to users in the form of an expandable and clickable tree. Such a tool could potentially aid users in finding what they need and help the NCHI staff to see a more logical pattern in topic-selection trends.

REFERENCES

Bobbitt-Cooke, M. (1999). Executive summary. Retrieved March 18, 2006, from http://www.healthycarolinians.org/2010objs/execsummry.htm

Campbell, R. J., & Nolfi, D. A. (2005). Teaching elderly adults to use the Internet to access health care information: before-after study. *Journal of medical Internet research electronic resource, 7*(2).

Eakin, D., Jackson, S. J., & Hannigan, G. G. (1980). Consumer Health Information - Libraries as Partners. *Bulletin of the Medical Library Association, 68*(2), 220-229.

Fox, S. (2005). *Health information online*. Washington, D.C.: Pew Internet & American Life Project.

Fox, S., & Rainie, L. (2000). *The online health care revolution: how the Web helps Americans take better care of themselves*. Washington, D.C.: Pew Internet & American Life Project.

Gollop, C. J. (1997). Health information-seeking behavior and older African American women. *Bulletin of the Medical Library Association, 85*(2), 141-146.

Greenberg, J. (2001). Optimal query expansion (QE) processing methods with semantically encoded structured thesauri terminology. *Journal of the American Society for Information Science and Technology, 52*(6), 487-498.

McGregor, B. (2005). Constructing a concise medical taxonomy. *Journal of the Medical Library Association, 93*(1), 121-123.

Patrick, T. B., Monga, H. K., Sievert, M. E., Houston Hall, J., & Longo, D. R. (2001). Evaluation of controlled vocabulary resources for development of a consumer entry vocabulary for diabetes. *Journal of medical Internet research electronic resource, 3*(3), E24.

Poensgen, A., & Larson, S. (2001). *Patients, physicians, and the Internet: myth, reality and implications*. Boston: Boston Consulting Group.

Rainie, L., & Horrigan, J. (2005). *Internet: the mainstreaming of online life*. Washington, D.C.: Pew Internet & American Life Project.

Shiri, A., & Revie, C. (2006). Query expansion behavior within a thesaurus-enhanced search environment: A user-centered evaluation. *Journal of the American Society for Information Science and Technology, 57*(4), 462-478.

Zeng, Q., Kogan, S., Ash, N., Greenes, R. A., & Boxwala, A. A. (2002). Characteristics of consumer terminology for health information retrieval. *Methods of Information in medicine, 41*(4), 289-298.