

Evaluation of Thesaurus-Enhanced Search Systems

For this assignment, in which we were asked to evaluate one commercial and one free thesaurus-enhanced search system, I chose to evaluate PsycINFO, a product of EBSCO Industries, Inc., and the California Environmental Information Catalog, a product of the State of California. In this paper I will evaluate the interfaces and retrieval of both systems, comparing the features of the two thesauri and the results of thesaurus-based searches in each of the two systems.

Evaluation of Interface

Both PsycINFO and the California Environmental Information Catalog are thesaurus-enhanced systems. Within each system, it is possible both to search and to browse within the thesaurus to select approved terms for use in searching the database itself. Various features are provided within each of the interfaces to help users perform thesaurus-based searches. On the landing page for the PsycINFO system, after entering through the link on the San Jose State University King Library website, the thesaurus is available through a prominent link in the uppermost horizontal navigation bar. Once the link to the thesaurus is clicked, the system navigates to a new page. It is easy to tell that one has entered the thesaurus, because the color scheme changes from primarily green to primarily blue and a label near the top of the page reads, “Browsing: PsycINFO—Thesaurus.” An alphabetical index, which provides both preferred and non-preferred entry points, is displayed below the thesaurus search box. Once the user has entered the thesaurus, he or she can browse the index or search for terms by entering them in the thesaurus search box. Beneath the thesaurus search box are three radio buttons: “Term Begins With,” “Term Contains,” and “Relevancy Ranked.” Performing a “Term Begins With” search

retrieves all terms beginning with the term entered; for example, the term “parental” retrieves “Parental Absence,” “Parental Attitudes,” “Parental Authoritarianism USE Parenting Style,” etc. Performing a “Term Contains” search retrieves all terms containing the search term either in the term itself or in the term’s scope note; for example, the term “parental” in this case retrieves “Animal Parental Behavior,” “Animal Rearing,” and “Authoritarianism (Parental) USE Parenting Style.” The “Relevancy Ranked” search function does not appear to work as well as the other two. Performing this type of search using the same term, “parental,” retrieves a list of terms not seemingly ranked by relevance but rather appearing in alphabetical order, although they do all appear to represent related concepts. Performing this type of search using the more specific term “parental absence” does retrieve that term first, but all subsequent terms appear alphabetically rather than by any true relevancy ranking. It is important to note that all preferred terms in the PsycINFO thesaurus index are hyperlinked to their own records, in which scope notes, history notes, standard relationships, and use for references are displayed. Broader, narrower, and related terms are displayed with checkboxes that can be checked off to add terms to a query using Boolean operators, and “Explode” and “Major Concept” checkboxes are included for each term as applicable. Records are displayed in table format, making them easy to read.

The ability to construct a database search using Boolean operators, the ability to “Explode” terms, and the ability to select a “Major Concept” for a term are additional features of the PsycINFO interface that aid users in performing thesaurus-based searches. To “explode” a term is, according to PsycINFO’s Help file, to “retrieve all references indexed to that term as well as all references indexed to any narrower subject terms”; selecting the “major concept” for a

term “[creates] a search query that finds only records for which the subject heading is a major point of the article” (EBSCO Industries, Inc., 2011a). Constructing Boolean queries involves checking boxes next to terms in the index, selecting the desired Boolean operator (OR, AND, or NOT) from a pull-down menu, and clicking the “Add” button. In this way I constructed the two queries I will describe later in this paper:

- (DE "Ability Level") AND (DE "Parent Child Relations")
- (DE "Aptitude Measures") AND ((DE "Ability Grouping") OR (DE "Educational Placement"))

I accessed the California Environmental Information Catalog (hereafter referred to as CEIC) via the URL provided on the List of Thesaurus-enhanced Search Systems which was part of the description of Assignment 3 for this course

<http://gis.ca.gov/catalog/thesaurus.epl?mode=browse>). Like PsycINFO, the CEIC interface includes various features to help users perform thesaurus-based searches. First, right above the thesaurus search box are three browsable, hyperlinked categories or facets: “Natural resources,” “Natural environment,” and “Human environment.” These represent the three fundamental facets of this thesaurus. Clicking on any of them redirects the user to a list of all the subfacets within that category. By clicking on different terms starting at the highest facets, the user can “drill down” through the thesaurus to its narrowest terms, and also find related terms.

Other features of the CEIC interface which help users perform thesaurus-based searches are the two buttons that appear below the thesaurus search box, “Enter Term” and “Similar Terms.” Performing an “Enter Term” search retrieves the term entered if it is a preferred term or a “Use” reference if it is a non-preferred term. For example, a search for “Agriculture,” a

preferred term, retrieves the record for “Agriculture,” whereas a search for “Husbandry,” a non-preferred term, retrieves a “Use Agriculture” reference, where “Agriculture” is hyperlinked to the record for that term. (A search for a term which is not contained as an entry point in the thesaurus, such as “shopping,” generates an error message: “Shopping is not in the...thesaurus.”)

Once the user accesses the record for a preferred term in the thesaurus, it is listed with its standard relationships. The CEIC thesaurus uses a two-column layout, with terms and relationships listed in the left-hand column and the search box and buttons in the right-hand column. The display is clean and easy to read, although the labels on the buttons are rather confusing; the distinction between “terms” and “themes” is not apparent, and there is no explanation of what the “Theme Filter” is or how to use it. Broader terms, narrower terms, related terms, and use for references are listed in the left-hand column under the term itself, and all preferred terms are listed with checkboxes, with their text hyperlinked to their own respective term displays. The checkboxes can be checked to construct queries, but unfortunately the system seems to be able to handle only simple searches on single terms, which I learned through trial and error. For example, in the record for “Wetlands” in the CEIC thesaurus, with no other terms checked off in the list of broader, narrower, or related terms, clicking the “Query on Selected Themes” button retrieves 688 matching records. However, checking off all terms in the list, including not just narrower terms but also broader and related terms, retrieves the same number of records. A similar query on the selected theme of “Marine resources” produces a similar result. The “Theme Filter” does not appear to work at all; any time I created a query by checking multiple terms and clicking the “Save to Theme Filter” or “Add to Theme Filter” button and then clicking the “Search” button on the subsequent screen, the system generated an “Internal Server

Error” and directed me to contact the server administrator (Figure 1). This occurred whether the boxes for “Location/Region: California,” “Themes/Keywords,” or “Data Types” were checked or not.

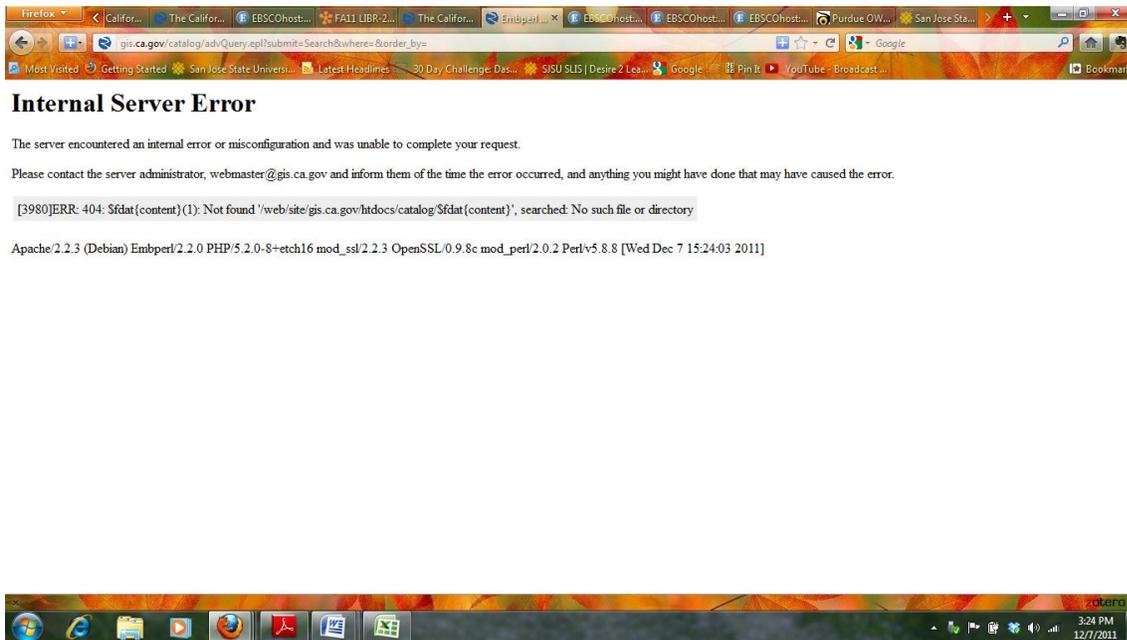


Figure 1. Internal Server Error generated by CEIC Theme Filter.

Performing a “Similar Terms” search in the CEIC thesaurus retrieves a hyperlinked list of terms that are related to the search term, but not necessarily the same list of “Related Terms” that is retrieved when you click on the “Enter Term” button. These terms seem to be related via a keyword index. For example, a “Similar Terms” search for “agriculture” retrieves a long list of terms that include some form of the word “agriculture” in them, whereas an “Enter Term” search for “agriculture” retrieves a much shorter list of related terms, many of which do not include any form of the word “agriculture” (such as “farming” and “irrigation”).

Besides the displays provided within each system's thesaurus, there are additional displays available. PsycINFO offers several displays. Search display provides four different search modes ("Boolean/Phrase," "Find all my search terms," "Find any of my search terms," and "SmartText Searching") as well as the ability to "apply related words" and to limit results by a variety of filters (e.g. publication name, full text, publication date range by year, publication type, language, methodology, etc.). In addition, the user can select either Advanced Search mode (the default), Basic Search mode, or Visual Search mode, and a Search History can also be displayed. Thesaurus display, described earlier, allows for either browsing or searching of the thesaurus. Cited References search display allows the user to retrieve "a list of citation records for the search terms you entered" (EBSCO Industries, Inc., 2011a). Indexes display allows the user to browse or search the various indexes within the database. The database can also be searched and displayed in a wide variety of languages and scripts, and a Help file (EBSCO Industries, Inc., 2011a) is included.

CEIC also offers several displays. Within the thesaurus, the default view is the search view, the features of which I have already described. As previously mentioned, it is also possible to browse the thesaurus starting with the three fundamental facets, and as one drills down through the thesaurus, individual terms are displayed in the same way as they would be if retrieved through a search. For terms which are indexed to resources in the catalogs contained in CEIC, clicking the "Query on Selected Themes" button generates a display of those resources in table format, with the particular catalog listed in the left-hand column and the particular resource listed in the right-hand column. An example of a term which is indexed to resources in the catalog is "Aquatic resources." The catalog and resource entries are hyperlinked to their

respective database records. The catalog records include catalog details, collaborating catalogs, and datasets, while the resource records include citation information, identification information, distribution information, and metadata information. As mentioned previously, the “Save as Theme Filter” and “Add to Theme Filter” buttons generate a simple table display with “On/Off” checkboxes in the left-hand column and “Filter Names” in the right-hand column, but unfortunately, utilizing the Theme Filter generates an Internal Server Error (Figure 1). Besides these thesaurus display features, navigating to other parts of the catalog via the “Discover” tab in the navigation bar near the top of the page leads to additional displays of the various catalogs that comprise CEIC.

When it comes to advanced features, PsycINFO differs from CEIC. When searching the PsycINFO database itself (not the thesaurus), if the user searches for a non-preferred term, records are still retrieved. There is no indication that the system has searched for alternative terms, and the user is not prompted to do so. A general keyword search, perhaps in coordination with the thesaurus, seems to be at work behind the scenes. Examining the subject descriptors associated with the retrieved records, the preferred term is listed for many (but not all) of them. For example, a search for “Parental Influence” (a non-descriptor) retrieves over 1200 records. Even on just the first page of results, several of the records have the preferred term, “Parent Child Relations,” listed as subject descriptors. Of those that don’t, many have the words “parental” and “influence” in the title of the work represented by the record.

Searching the CEIC catalog itself (not the thesaurus) using non-preferred terms leads to some interesting results. For example, searching for “husbandry,” a non-preferred term in the thesaurus, retrieves one record. Clicking on the record reveals that the word “husbandry”

appears in the Description field. On the other hand, searching for “agriculture,” the preferred term for “husbandry,” retrieves a great many records—and the one retrieved by the search for “husbandry” is not among them. As with PsycINFO, there is no indication that the system has searched for alternative terms when non-preferred terms have been entered, and the user is not prompted to do so. This section of the CEIC website also features an “Advanced Search” tool. Clicking the “Advanced Search” link after entering the catalog via the “Discover” tab brings up a new search interface which includes filters for information types, subjects, location, title, and “Matching Cataloged Resources.”

The advanced features on the CEIC website are certainly nice, but my examination of them has raised some serious usability concerns. First, the architecture of the site is somewhat confusing. It is hard to tell where the boundaries lie between the thesaurus, the catalog, and the website on which they reside, so navigating the site is difficult. After clicking on links from within the thesaurus, it is possible to exit the thesaurus itself without knowing that this is happening. The content of the site is not very well organized overall. Even the file structure of the site, as indicated in the URLs of various pages, is confusing, as the subdomains of ca.gov change at the click of a link. Also, the labels on some of the buttons do not clearly indicate what functions they perform (e.g. “Query on Selected Themes,” “Save as Theme Filter,” and “Add to Theme Filter.”) As I previously mentioned, I did not find any place where the Theme Filter was described or explained, nor did I find any “Help” file within the thesaurus or the larger database, so I was left to my own trial and error to learn how to use this system effectively. In spite of these many usability issues, however, the thesaurus itself is well-designed, and the browse feature is particularly useful and user-friendly.

In contrast, the PsycINFO interface is very user-friendly overall. The interface is simple and clean in appearance and well-organized. In her book *Designing Interfaces*, interaction designer Jennifer Tidwell presents a variety of ideas and patterns that can be employed to solve common design problems and create usable interfaces. The PsycINFO system employs many of them. Some usability techniques employed in this interface include what Tidwell describes as “extras on demand,” which “show the most important content up front, but hide the rest, [letting] the user reach it via a single, simple gesture” (p. v), “multi-level help,” which provides a variety of different help techniques (p. v), a few task-oriented and descriptive “clear entry points” (p. v), “color-coded sections,” which indicate groupings or sections of pages (p. vi), “button groups” to present related actions (p. vii), and “contrasting font weights,” which “separate different levels of information and add visual interest (p. ix). Overall, the PsycINFO interface is easy to understand and to navigate, giving it a high rating for usability.

Evaluation of Retrieval

Both PsycINFO and CEIC allow users to construct search queries by selecting preferred terms from their thesauri. As I mentioned earlier, I performed two thesaurus-based searches in PsycINFO:

- (DE "Ability Level") AND (DE "Parent Child Relations")
- (DE "Aptitude Measures") AND ((DE "Ability Grouping") OR (DE "Educational Placement"))

My first query, for “Ability Level” AND “Parent Child Relations,” retrieved nine results. As expected, the records retrieved included both terms as subject descriptors. While it is possible that some or all of the same words could have been used in other indexes to index additional

documents in the database, these nine documents were the only ones indexed with both of these terms as subject descriptors. In list display, not all of the records showed both descriptors in the display, but this was revealed by clicking through to the detailed record for individual items. While thesaurus terms found in the result set were not highlighted, they did appear as hyperlinks. Clicking on any particular hyperlinked descriptor initiated a new search using just that term. This appeared to be the only way to browse thesaurus terms from the retrieved results without exiting to the thesaurus itself.

By changing one of the preferred terms in my search query, “Parent Child Relations,” to a non-preferred term, “Parental Influence,” I was able to observe how the system handles searches containing terms not found in the thesaurus. While my initial search retrieved nine records, my new search retrieved 20,019 records! A message appeared at the top of the list of retrieved records: “Note: Your initial search query did not yield any results. However, using SmartText Searching, results were found based on your keywords.” Based on this message and my examination of some of the retrieved records, it appears that in this case, the system reverted to an exclusively keyword search and ignored the subject index entirely, for at least some of the records retrieved did not include as a subject descriptor the one preferred term, “Ability Level,” that was still included in my query.

My second query, for ("Aptitude Measures") AND (("Ability Grouping") OR ("Educational Placement")), retrieved 27 results. As expected, the records retrieved included the term “Aptitude Measures” in combination with either “Ability Grouping” or “Educational Placement” as subject descriptors. My observations of the system’s performance with this query were consistent with those of my first query.

In CEIC, the results observed by searching for documents indexed with individual terms in the thesaurus are inconsistent with those observed by combining terms in the default search field accessible via the “Discover” tab from the thesaurus. In my first query, I first searched for the individual terms “Forest management” and “Water conservation” in the thesaurus, then clicked the “Query on Selected Themes” button from within their respective term records. The resulting display indicated that 15 records were retrieved for “Forest management” and 32 for “Water conservation.” I deduced that there should be a maximum of 15 records retrieved when I performed a search for (“forest management”) AND (“water conservation”) in the search field for the database itself, since “Forest management” would have to be indexed to every record. However, I would estimate that more than six times that number of records was retrieved in this instance. (I say estimate because the system does not indicate the total number of records retrieved in this display.) Clearly, a different type of search is occurring in these two examples.

For my second query of CEIC, I decided to use the same terms, “Forest management” and “Water conservation,” but this time I tried using the Boolean operator OR rather than AND. I deduced that in this case, there should be a maximum of 47 records retrieved, since at least one of the terms would have to be indexed to every record, and assuming that they never appeared together, the maximum number of records that could be represented by one or the other term would be 47. However, when I performed my search, I got what appeared to be the same results as I got in my first search, when I used the AND operator. I concluded from these two searches that the CEIC system does not permit Boolean searching, at least within the larger database. And as previously mentioned, I was unable to construct complex queries within the thesaurus due to the “Internal Server Error” (Figure 1) that resulted every time I attempted to do so.

The question remains, then, how was the system executing searches on my queries in these two examples? Judging by the list of retrieved records, it was performing simple keyword searches. In the list of results for my second query, the individual words “management,” “conservation,” “water” and “forest” (and other forms of those words) appeared in boldface. Returning to my earlier example, a search of the system for a non-preferred term such as “husbandry” suggests that the thesaurus is not actually employed in searches performed in this section of the site, as records retrieved via that search term differ from those retrieved via the preferred term, “agriculture.” This supports my conclusion that the system is performing only a keyword search and not a thesaurus-based search here. For this reason, some of the considerations that applied to my examination of PsycINFO’s performance as a retrieval system, such as how the interface shows results from the thesaurus, whether it highlights thesaurus terms found in the result set, and whether the interface allows users to browse thesaurus terms from the retrieved results, do not apply here. Since the system does not appear to be coordinating its search with the thesaurus, it does not provide feedback on terms not found in the thesaurus, and there is no non-descriptor to descriptor redirection. Finally, as previously mentioned, in this display, the interface does not indicate the total number of records retrieved by a particular search.

Usability Critique and Suggestions for Improvement

The PsycINFO and CEIC systems both have strengths and weaknesses. From a usability perspective, I think PsycINFO definitely wins out over CEIC, for the reasons mentioned in my interface evaluation above: the PsycINFO interface is simple and clean in appearance, well-organized, and easy to navigate, while the CEIC interface is not. PsycINFO’s inclusion of a

Help file is particularly valuable. The design of the site makes for an efficient experience by the user, minimizing time spent backtracking and recovering from errors. The interface is easy to learn and highly readable. Since this interface already scores high in usability, it is difficult to identify suggestions for improvement, but a few can be made. One suggestion is to include a “Thesaurus Terms A to Z” index that persists throughout all displays within the thesaurus, allowing the user to click on any letter and immediately jump to that section of the index. Another suggestion is to improve the “Relevancy Ranked” search function so that its rankings are truly relevant. Finally, a minor suggestion is to fix the alignment of the checkboxes in the “Explode” and “Major Concept” columns within term records. At least in my browser (Firefox), they don’t line up properly, which may be confusing to some users.

The CEIC interface, as previously discussed, has several major usability issues. The lack of a Help file is a major oversight, as users do not have any documentation to turn to when experiencing difficulties using the system. Ambiguity in labels (e.g. the use of “term” and “theme” on buttons within the thesaurus, with no disambiguation) leads to user errors and inefficient squandering of the user’s time. The lack of clear distinctions between the thesaurus, the database, and the larger website leads to confusion on the part of users who may inadvertently wander into the wrong area of the site. And the Internal Server Error consistently generated by the “Theme Filter” (Figure 1) is a major obstacle for users wishing to construct complex queries. Because of these many usability issues, there are many areas in which CEIC can be improved. The addition of a Help file that defines terms, provides explanations of different features, and anticipates frequently asked questions would go a long way in assisting users who run into challenges. Easily confused verbiage such as “term” and “theme” should be

explicitly defined. Users should be provided with a roadmap that clearly indicates where they are (i.e. thesaurus, database, or website) within the system at any given time. This could be accomplished through the use of varied color schemes and simple labels as are employed in PsycINFO. Finally, the database or server administrator should test and repair the “Theme Filter” or remove it from the site entirely.

Overall, I found this assignment to be a very effective learning experience. Examining these two search systems from both an interface and a retrieval perspective revealed strengths and weaknesses in both and made me more aware of the design and architecture of such systems. This exercise helped me to understand what does and does not work with regard to interface design and thesaurus-based searching. I know that the knowledge I have gained through this exercise will serve me well as I move forward in my studies and career as an information professional.

Bibliography

- EBSCO Industries, Inc. (2011a). EBSCO Help. In PsycINFO [database]. Retrieved from http://support.ebsco.com.libaccess.sjlibrary.org/help/index.php?help_id=606, December 7, 2011.
- EBSCO Industries, Inc. (2011b). PsycINFO [database]. Retrieved from <http://web.ebscohost.com.libaccess.sjlibrary.org/ehost/search/advanced?sid=d836b6b0-1005-43f4-94e7-f2cca69d1f7e%40sessionmgr4&vid=7&hid=12>, December 7, 2011.
- State of California. (2007). California Environmental Information Catalog. Retrieved from <http://gis.ca.gov/catalog/thesaurus.epl?mode=browse>, December 7, 2011.
- Tidwell, J. (2006). *Designing interfaces*. Sebastopol, CA: O'Reilly Media, Inc.