SciFinder Scholar 2006: An Empirical Analysis of Research Topic Query Processing

A. Ben Wagner*

Science and Engineering Library, Arts and Sciences Libraries, University at Buffalo, 226 Capen Hall, Buffalo, New York 14260-1672

Received November 1, 2005

Topical search queries in SciFinder Scholar are processed through an extensive set of natural language processing algorithms that greatly enhance the relevance and comprehensiveness of the search results. Little detailed documentation on these algorithms has been published. However, a careful examination of the highlighted hit terms coupled with a comparison of results from small variations in query language reveal much additional, useful information about these algorithms. An understanding of how these algorithms work can lead to better search results and explain many unexpected results, including differing hit counts for singular versus plural query words and phrases.

I. INTRODUCTION

This article examines in detail the natural language query (NLQ) processing algorithms involved in the SciFinder Scholar (SFS) subject search feature, *Research Topic*. The SFS search system was developed by Chemical Abstracts Service (CAS) and has been marketed to academic institutions since 1998.[1] The commercial version, simply named SciFinder, was introduced three years earlier in 1995.[2] Both versions appear to use the same algorithms in processing topical search queries. SciFinder Scholar 2006, the current academic version, was released in August 2005. A detailed time line of the development of SciFinder Scholar appears in the appendix. The Research Topic icon is reached via the Explore icon on the SFS main screen. General aspects of this NLQ processing are described in the SFS User Guide[3] including the use of prepositions, Boolean operators, abbreviations, and automatic truncation. In general, these same features are also apparent by actual use of the system. However, there are many subtle, undocumented aspects of the query processing. Understandably, CAS has been careful not to reveal too many details of their algorithms, given the highly competitive nature of database search interfaces. Nearly all of the information in this paper has been derived by the input of slightly varying queries and close examination of the hit terms highlighted in the results display. Because words causing a given record to be retrieved are highlighted in blue in the results display, an empirical analysis of those results can reveal much about these algorithms. But first, a brief introduction to the SciFinder Scholar retrieval system is in order.

II. BACKGROUND AND LITERATURE REVIEW

SFS is an elegant search interface to six core chemical-related databases. Five of these databases are produced by CAS itself:

- CAplus, the main Chemical Abstracts literature database of over 23 million references. CAplus includes in-process records, over 21 000 pre-1907 references, and document types such as letters to the editor that do not appear in print Chemical Abstracts.
- CAS REGISTRY, the substance database of over 27 million organic and inorganic substances and 57 million biosequences.
- CASREACT, providing access to over 10 million reactions from the journal literature and patents.
- CHEMLIST, the regulatory chemicals database of over 238 640 compounds compiled from an extensive group of state, national, and international regulatory lists and inventories.
- CHEMCATS, listing suppliers of commercially available chemicals worldwide.

The sixth file, added in 1999, is the complete National Library of Medicine's (NLM) MEDLINE database back to 1950 containing over 15 million references. MEDLINE includes all but approximately 1% of the citations in PubMed, NLM's freely available Web database. All six files are also searchable via a number of Web interfaces on the STN International system. STN is a cooperative effort of CAS (U. S. A.), FIZ Karlsruhe (Germany), and the Japan Science and Technology Corporation.

SFS has many excellent features including:

- Easy-to-use links between compound, reaction, regulatory, commercial supplier, and literature information
- Facile access to electronic full-text versions of references via ChemPort, an Internet-based linking system
- Substructure searching of reactions and substances
- A large and growing set of experimental and calculated physical property values
- Extensive coverage of life sciences, physics, materials science, and engineering

SciFinder Scholar was designed to be an intuitive system, requiring little training to quickly produce relevant results. On the whole, the system meets that objective.

Ridley[4] has published a fine book reviewing all aspects of the system. Other good overviews include works by Nitsche and Buntrock,[5] Schwall and Zielenbach,[6] and Haldeman et al.[7] A number of specific aspects have been

* Author phone: (716)645-2947 x230; e-mail: abwagner@buffalo.edu.

10.1021/ci050481b CCC: $33.50
Published on Web 02/24/2006

studied including a comparison of citation searches between SciFinder Scholar and Web of Science,[8] its use in undergraduate education,[9−11] reaction searching,[12,13] and substructure searching.[14,15] However, this article appears to be the first to extensively analyze the Research Topic feature.

The rest of this article focuses on the query processing behind the Research Topic feature, which searches CAplus and MEDLINE databases. All retrieval counts are from searches performed since September 26, 2005. Actual keyword query statements are placed in single quotes.

III. GENERAL CHARACTERISTICS OF *RESEARCH TOPIC* QUERIES

Generally, the best results are obtained by liberal use of prepositions, for example, 'use of antibiotics to prevent diseases in cows'. SFS recognizes prepositions but does not analyze their linguistic meaning. They are used only to break the query into component concepts. Hence, the common Internet search engine strategy of typing in a string of keywords separated by spaces usually leads to suboptimal results.

No truncation symbols are recognized because the NLQ processing provides automatic truncation. Boolean operators are usually unnecessary and inadvisable. SciFinder Scholar developers[16] indicate that "AND", "OR", and "NOT" are not treated as Boolean operators, but rather as conversational English. However, once the NLQ algorithm identifies query concepts and performs the search, Boolean logic is used to generate the pick list of the various search results presented to the user, as will be seen in section IV. Because the English and formal Boolean senses of the word "NOT" are the same, using "NOT" in a query will effectively eliminate false hits which can be readily identified by one or two keywords.

Parentheses can only be used to specify up to three synonyms separated by commas after a given term. An example of this syntax would be 'paints (coatings) which are fire retardant (flame resistant, intumescent)'.

Case is ignored. Searching 'AIDS' or 'aids' generates identical results. Also ignored are single and double quotes. Given the size and technical nature of the database, query words are not run against a spelling checker, except for some of the most common misspellings.

The general description in the preceding paragraphs is confirmed in the SFS User Guide.[3] All of this may sound rather limiting to power searchers who often use long, precisely structured queries. However, the goal of SciFinder Scholar is to permit the effective exploration of topics without extensive training or special syntax. CAS never intended SFS to be a wholesale replacement for other STN interfaces such as STN Easy or STN Express.

The SFS NLQ processing system uses a number of techniques including synonym generation and a combination of query concepts in different ways to provide useful results. Like the best Web search engines, it has a remarkable ability to find good hits on almost any topic remotely related to chemistry. When teaching SFS to students, the author makes this point by retrieving excellent hits on using lasers to clean paintings, a topic far removed from pure chemistry.

SciFinder Scholar Research Topic's strength is relevance not in every case comprehensiveness. If one asks a general question, one gets very good, but general, answers. A more specific question gives a more specific and often a more comprehensive set within the bounds of the stated topic. One example is that the general query 'recovery of precious metals' retrieves 1954 hits where the two concepts are closely associated within each record. The more specific query of 'leaching of gold' retrieves 3778 hits for closely associated concepts. It should be noted that Chemical Abstracts indexes articles as specifically as they are written and does not use a subject heading tree structure with an Explode feature like MEDLINE.

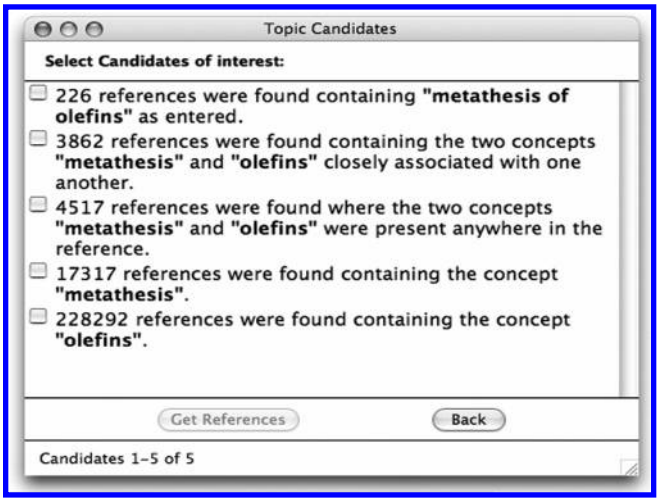

Figure 1. The SciFinder Scholar Topic Candidates dialogue box. Query statement: "metathesis of olefins".

IV. PRESENTATION OF ANSWER SETS OPTIONS OVERVIEW

Before proceeding to a more detailed review of SFS query processing, it is helpful to understand how answer sets are presented to the user once the query has been processed and run against the database. Rather than immediately displaying results, SFS presents a number of options as a pick list called the Topics Candidates dialogue box. As illustrated in Figure 1, each item on the pick list represents a different combination of the component concepts as determined by the NLQ algorithms.

For a three-concept search (A, B, and C), the Topic Candidates dialogue box presents every permutation: A+B+C, A+B, B+C, A+C, and each individual concept. In addition, four categories of results are usually presented:

- *As Entered*: essentially an exact phrase option as input by the user with little or no modification by the NLQ algorithms.
- *Concept* (single word or phrase): makes use of autotruncation, synonyms, and other algorithms to expand the query to closely related terms. This is discussed in depth in the next section.
- *Closely Associated with One Another*: the listed search concepts are in the same sentence (title and abstract) or within the same index entry.
- *Present Anywhere in the Reference*: equivalent to a simple Boolean "AND" of the listed concepts.

Hence, it is possible for each query to generate a number of topic candidate sets. The user can select as many of these sets as desired. At first glance, the amount of verbiage in the Topic Candidates dialogue box can appear intimidating. However, the options are presented in a consistent order (*As*

Entered is always first, if present), show the number of answers in the set, and are clearly worded. Many examples of the differences among these various answer sets will be discussed in the rest of this paper.

V. DETAILS OF THE NATURAL LANGUAGE QUERY PROCESSING

Some aspects of the query processing are obvious by examining the highlighted terms in a retrieval set, such as automatic truncation and pulling in synonyms from a dictionary. Other aspects, especially the order in which various algorithms are applied, are difficult to ascertain.

The only information found on the sequence of operation was published by Williams[2] of the Monsanto Company. Williams gives this sequence: (1) removing case, (2) breaking the query into concepts by removing prepositions and stop words, (3) checking an internal dictionary of synonyms, alternate spelling forms, and abbreviations, (4) checking the CAS Registry File of chemical substances and adding the CAS Registry Number to the query if an exact name match is found, and (5) truncating any remaining terms.

Overall, this description appears reasonable, but there is far more happening in the current version than Williams' brief description reveals. Though most queries appear to run through the entire sequence of algorithms, there is one known situation where the processing of query terms ceases before the autotruncation step. This case, which affects the equivalence of singular versus plural query statements, is discussed in section VII.

V.A. Parsing the Query—Identifying Concepts. As previously mentioned, case is ignored. Single quotes, double quotes, and periods are replaced with spaces. A comma between two terms is usually interpreted as a Boolean "OR" and prevents the two terms from being searched as a phrase. The use of other punctuation such as the pound sign (#) and exclamation point (!) produces unpredictable results. Sometimes, the punctuation is retained in the search term, and this typically generates few, if any hits. Other times, the punctuation appears to be stripped away at this stage or when autotruncation occurs near the end of the processing.

It appears that the next step is to separate the query into its component concepts or phrases using prepositions, other stop words, and conjunctions as break points. The linguistic meaning of prepositions is not analyzed. Once the concepts are defined, the prepositions are discarded. The queries 'Analysis in bromine' and 'Analysis of bromine' both generate the same result, a two-concept search of 'analysis' and 'bromine'. Prepositions have a minor effect on exact phrase searching, but this will be discussed later.

Stop words are common, highly posted terms which have no search value because of their frequency or generic nature. "Effect", "study", and "search" are stop words as well as the expected articles like "an" or "the". The only way to discover that a stop word has been eliminated is to notice its absence from the concepts listed in the Topics Candidate dialogue box. Some words such as "using" that might be expected to be a stop word are not.

Long complex query statements should be avoided, especially because Boolean operators are not processed as such and parentheses cannot be used to group terms. It is nearly always preferable to initially input a simple query and then use Analyze/Refine options to narrow the results. The use of Boolean operators can defeat some of the effectiveness of the NLQ system.

As noted in the SFS User Guide,[3] another problem in attempting to use Boolean statements is that the NLQ system does not usually distribute modifiers. The query 'liver or ovarian cancer' will retrieve all records containing the concept "liver" OR the concept "ovarian cancer" rather than the likely intended "liver cancer" OR "ovarian cancer". However, the system does handle certain queries as presumably intended by user. The query 'effect of DDT on birds and fish' and the query 'effect of DDT on birds or fish' both give the same result, "DDT and (birds or fish)".

V.B. Generating Synonyms and Alternate Forms—the "Secret" Dictionary. SciFinder Scholar analyzes the resulting phrases and terms from the parsing operation against a dictionary of synonyms, abbreviations, acronyms, and alternate word forms. All generated alternative terms are then used in performing the actual search. For competitive reasons, the dictionary/thesaurus cannot be accessed directly. However, given that all of the terms that cause the retrieval of a given record are highlighted in blue, one can surmise a fair amount about this "invisible" dictionary. The recognition of abbreviations, common misspellings, and American/British spellings are noted in the SFS User Guide.[3]

CAS standard abbreviations and acronyms (http://www.cas.org/ONLINE/standards.html) such as NMR are always recognized and expanded to the full term and vice versa. Many nonstandard abbreviations are recognized as well. BTU is equivalent to British Thermal Unit as is XRD for X-ray diffraction. Some reasonably common abbreviations are not recognized. GPS is not recognized as global positioning system, nor is FTIR expanded to the full term. EPA does not retrieve 'USEPA' or 'Environmental Protection Agency'. Of course, many acronyms have multiple meanings, making fixed equivalents problematic. The only way to determine if matches have been made against the synonym dictionary is by inspecting the search results.

In some cases, additional synonyms are added to the full-term−abbreviation pair. Prep, prepn, preparation, synthesis, and syntheses are all equivalents to each other. The use of any one of those terms results in the retrieval of all of them. Chemical names that are acronyms are handled by the Registry database name matching algorithm that is discussed in the next subsection.

Alternate word forms are often recognized. Freeze, frozen, and freezing are all equivalent. Bleed retrieves bled and blood, which may not always be a good thing. Inconsistencies are not hard to find. Broken does not retrieve break, breaks, or breaking. Blow retrieves blown, probably because of autotruncation, but not blew.

Regular singular−plural forms presumably are handled by the autotruncation algorithms. However, at a *Concept* results level, singular and plural query words and phrases often generate differing hit counts. This unexpected finding is discussed in section VII. Many irregular plurals are correctly handled, either via autotruncation or specific dictionary entries: butterfly and butterflies; mouse and mice; flamingo, flamingoes, and flamingos. The system makes rat equivalent to rats while excluding rate. Interestingly, it does not seem to map many "f" to "v" or "um" to "a" transformations. Hoof

is equivalent to hoofs and hoofed but not hooves. Wolf does not retrieve wolves, and compendium does not retrieve compendia.

Certain terms are extensively linked to synonyms. Oxen maps to cattle, calves, bovine, cows, and bull (unfortunately including bull trout records!). Human(s) is equivalent to patient(s), person(s), man, men, woman, and women. Cancer generates neoplasm, carcinoma, tumor, tumour, and carcinosarcoma. HIV is equivalent to AIDS, acquired immunodeficiency syndrome, and human immunodeficiency virus. This is an example of the crossover between synonym generation and acronym recognition.

Most equivalents seem to go both directions; that is, inputting either term gives the same results. A few, however, work only one way. 'Dairy products' retrieves 'milk', but 'milk' retrieves neither dairy nor dairy products.

Spacing and irregular transformations are often not recognized. Long is not equivalent to length, nor is lakewater to lake water. However, 'ground water' and 'groundwater' are equivalent. 'HIV1' and 'HIV 1' are not equivalent. 'Long horned beetle' does not pick up 'Longhorned beetle'.

The dictionary does not work like a true thesaurus with an Explode feature whereby narrower terms are automatically included. In fact, broader or narrower terms are seldom automatically added. Pesticide does retrieve insecticide, acaridide, and nematocide but not parasiticide. Citrus fruit does not generate any mapping to specific citrus fruit such as oranges or lemons.

Equivalents exist for British−American spellings. So, color maps to colour, and aluminium sulphate will retrieve aluminum sulfate. There is no overall spell checking. However, some of the most common errors have dictionary equivalents. Hence, 'commercal' retrieves commercial. There is no warning that the correct spelling has been added to a query containing a misspelled term. Many typos are not picked up. 'Molydbenum' or 'molybdnum' do not map to the correct spelling, molybdenum.

Overall, it appears that a good portion of the dictionary has been built on an example-by-example basis, rather than by writing broad rules or using an extensive, existing thesaurus structure. CAS makes changes and additions to this dictionary based on their own testing and user comments, at times even indirectly. This author publicly pointed out at a conference that 'dairy products' did not map to milk, and within a few months, that equivalence was added. Indeed, after the publication of this article, adjustments to the SciFinder dictionary/thesaurus may eliminate some of the examples cited in this paper.

The dictionary contains at least some of the equivalents found in *See* and *See Also* references in the Chemical Abstracts Collective Index Guides. The CAS 13th Collective Index Guide (1992−1996) provides a See reference from 'Northern white cedar' to the scientific name '*Thuja occidentalis*', and these are equivalents in SciFinder searching. 'Synthetic sponges' is equivalent to 'Sponges (artificial)', a See reference in an older Index Guide.

Selecting the *As Entered* answer set at the top of the Topics Candidate pick list completely eliminates any results generated by the dictionary. No synonyms or alternate forms including simple plurals are included in the *As Entered* results. This can be useful when unwanted synonyms or truncation takes place in the more general *Concept* retrieval sets.

V.C. Identifying Chemical Names−Adding Registry Numbers. The primary way chemical substances are indexed in Chemical Abstracts is by use of the CAS Registry Number (CASRN) system. This system, started in 1965, assigns a unique, hyphenated number to every compound indexed by CAS. One of SFS's most helpful algorithms matches the terms from the parsing algorithms with chemical synonyms in the master Registry File. The Registry File is the database of all assigned registry numbers with the official index names, synonyms, molecular formula, and much other information.

Each concept, whether a single term or phrase, is run against the millions of synonyms in the Registry File. If an exact match is found, the registry number is added to the search query. The match must be exact. Williams[2] notes that 'Nylon' does not match any synonym exactly, so no registry number is pulled into the search strategy. However, a specific nylon, such as 'Nylon 66', is an exact match whose registry number is then added. This addition of the registry numbers greatly enhances retrieval. Note that only the registry number is added to the query, not all of the other synonyms in the file. This is because there can be dozens or hundreds of synonyms, some of which are not unique to that compound. Through correspondence with a senior SciFinder programmer,[17] it was learned that, if an exact match is made on any synonym in the Registry File, no autotruncation is performed. This special case is discussed in section VII.

Even though stop words such as the article 'a' are normally removed, the synonym matching algorithm is sophisticated enough to match 'Vitamin A' with its registry number. Note that the registry number−name matching algorithm is bypassed when the *As Entered* answer set is selected, retrieving only records with exactly that synonym.

A registry number is allowed as a search term in Research Topic, though it usually is preferable to search the registry number in the Locate Substances: Substance Identifier search option.

There are cases where a simple, unique synonym searched in Research Topic can retrieve more hits than retrieving a substance record and getting the associated literature references. For example, the pesticide 'Mirex' generates 2506 answers, if the *Concept* option is chosen (which includes name and registry number searching). A registry number search provides for only 2253 hits because there are 253 Mirex records where Mirex is in the title or abstract but the registry number for Mirex was not assigned. Of the 253 Mirex records without registry numbers, 144 came from MEDLINE and 109 came from CAplus. Searchers can readily check the comprehensiveness of registry number assignment for uniquely named compounds by doing the following Research Topic query, 'substance name' NOT 'registry number'.

There is one case where registry numbers should be searched in Research Topic. Registry numbers with letter suffixes (D for derivatives and DP for derivatives preparation) should be searched using the Research Topic query box because one cannot limit to these suffixes when crossing over from a substance record to the literature references.

Searching for the "P" suffix (preparation) of registry numbers should be done by retrieving a substance record, clicking on the *Get References* button, and then choosing *References associated with Preparation*.

V.D. Autotruncation. As noted in the SFS User Guide,[3] SciFinder Scholar nearly always does automatic truncation. In cases where the truncation is more excessive than desired, four aspects of SFS help eliminate irrelevant citations:

- The ease with which search results can be analyzed and refined (limited)
- The ability to add additional concepts to eliminate false drops caused by the inappropriate truncation of one particular term
- Highlighting in the displayed records shows clearly where truncation was done
- The option to choose the *As Entered* hit set, which is an exact match on what the user entered; no truncation is done on those results.

Still, examples of inappropriate truncation are not difficult to find: mine also retrieves mineral and minimize; disposable also retrieves disposal, and diapers retrieves diaphragm(s), leading to a very noisy set for 'diapers which are disposable'; medical truncates at 'med...', retrieving medium, media, medium, medicine(s), and mediated; workers as a concept truncates at 'work...'; and tropical plants as a concept retrieves tropanes (a plant alkaloid).

V.E. Unintended Interactions between Algorithms. Chemical Abstracts Service has developed an effective natural language query processing system. The preceding discussion makes clear the benefits of carefully examining highlighted words in search results to gain insight into how the query was processed.

Occasionally, the various algorithms can interfere with each other. Searching the acronym 'SIDS' does helpfully expand to include Sudden Infant Death Syndrome. Unfortunately, the autotruncation algorithm truncates SIDS at the 'D', which pulls in side, sides, and so forth. One has to choose the *As Entered* option or Refine by Research Topic using terms such as 'infant or baby'. However, then, one may not retrieve references with only the acronym or only the full term.

VI. DETAILS OF MULTIWORD PHRASE PROCESSING

This section discusses how SciFinder Scholar handles multiword phrases. One caution should be made regarding SciFinder Scholar's hit term highlighting in records. All of the individual words contained in a phrase are highlighted, even when the record is actually retrieved on the basis of a specific relationship between those words. For example, a record retrieved on the basis of containing the exact phrase 'carbon nanotubes' will also have every isolated occurrence of 'carbon' and 'nanotubes' highlighted.

VI.A. *As Entered* Results for Phrases. Nearly all simple phrases will generate *As Entered* as the first option in the Topic Candidates dialogue box unless the exact phrase does not exist in the database. The *As Entered* option is word-order-specific and does not convert singular to plural forms or vice versa. As mentioned previously, none of NLQ processing algorithms described in section V are performed.

Interestingly, the system treats a preposition in a phrase as a place holder for any other word, doing the equivalent of a "within one word" search operation. The 'thin film semiconductors' *As Entered* set retrieves 175 hits containing that exact phrase. However, the 'thin film in semiconductors' *As Entered* set retrieves 222 hits containing the exact phrase plus records with phrases such as "thin film *device* semiconductors".

Table 1. Differences in Query Phrasing

query	as entered hits	concept hits
quantum size effect	3442	4814
effect of quantum size	180	10 573

VI.B. *Concept* Results for Phrases. Phrases with embedded prepositions are broken apart at the prepositions with each resulting word or phrase treated as a separate concept and processed through the full NLQ process described in section V. Various combinations of the concepts are presented in the Topic Candidates dialogue box.

Long phrases without embedded prepositions are generally broken apart into individual words, which are run through the full NLQ processing with the resulting concepts simply "ANDed" together. In other words, each record contains at least one term derived from every word in the query phrase. This provides appropriate retrieval for those accustomed to general Web search engines where strings of randomly ordered keywords are input with just spaces between words. However, the use of prepositions in natural phrases is still the best approach.

Given the complex, interacting algorithms used to process queries, slight differences in phrasing can produce dramatically different results, as seen in Table 1.

The small size of the second query's *As Entered* results is due to the tight word-order restriction. The large size of the second query's *Concept* results is due to the elimination of 'effect' as a stop word, giving 10 573 records on simply "quantum size". Interestingly, the first query's *Concept* results do not remove 'effect' as a stop word.

VII. SINGULAR VERSUS PLURAL QUERIES: AN UNEXPECTED DIFFERENCE IN HIT COUNTS

Given the extensive autotruncation that typically occurs, the author was surprised to find many examples where single versus plural words and phrases gave differing hit counts as *Concepts* in the Topic Candidates listing of search results. Different hit counts for the *As Entered* listings for singular versus plural forms are expected, because the results are essentially the exact phrase with no autotruncation. Though many words and phrases do provide identical *Concept* results for the corresponding singular and plural queries, three cases where this is not true were discovered.

Case 1. Match with a CAS Registry Number (CASRN) Synonym. As pointed out in section V.C, terms that exactly match a synonym in the CAS Registry File exit the NLQ algorithms before any autotruncation occurs. This occurs even for terms that happen to match a name associated with a generic substance with zero literature references, such as *carbon fibers*, CASRN 308063-56-1. The practical effect of this is seen in Table 2, where 'carbon fiber' gives very different results from 'carbon fibers'.

Because 'carbon fibers' matches a registry record, the *Concept* results include the 57 673 records with the exact phrase 'carbon fibers' plus the zero literature references

Table 2. CASRN Synonym Matching

query	as entered hits	concept hits
carbon fibers	57 673	57 673
carbon fiber	41 801	75 225

Table 3. Biological Organisms Singular and Plural Forms

query	as entered hits	concept hits
bald eagle	158	262
bald eagles	157	239

Table 4. Multiword Phrases Concept Hit Counts

query	singular concept hits	plural concept hits	form of general subject heading
photographic paper(s)	10 322	10 322	singular
nuclear wave function(s)	23 005	6848	singular
softening agent(s)	10 682	10 682	plural
soil amendment(s)	13 975	27 212	plural
chemical warfare agent(s)	5796	7813	plural
sputtering target(s)	17 093	20 527	plural
piston ring(s)	3048	3052	plural
candy bar(s)	213	455	not a heading

associated with registry number 308063-56-1. Compared to the 'carbon fibers' results, the 'carbon fiber' *As Entered* set is smaller because it will not retrieve records containing the plural phrase. The 'carbon fiber' *Concept* set is larger because it is not an exact match against a chemical synonym and, therefore, is run through the full NLQ query processing. Other examples of matches with Registry file synonyms include kaolin, tall oil, carbon nanotube, and liquid rosin. In each case, the *As Entered* set is identical to the *Concept* set, while kaolins, tall oils, carbon nanotubes, and liquid rosins give very different hit counts.

Case 2. Biological Organisms: Common and Scientific Names. As mentioned in section V.B., the SciFinder Scholar dictionary contains many equivalents between common and scientific names at both a genus and species level. For all queries tested using a common name in singular form for a biological organism, the *Concept* results included records with both the singular and plural forms of the name plus records with the scientific name highlighted as a hit term. However, queries using the plural form of the common name included only records with the singular and plural forms of the name but not those with only the scientific name. This is demonstrated by the bald eagle example in Table 3.

Other examples of this case can be tested by searching for osprey(s), daffodil(s), and sea lamprey(s).

Case 3. Certain Multiword Phrases. Other than the two cases described above, no cases could be found where a one-word term generated different results at a *Concept* level between its singular and plural forms. Laptop, window, veneer, calendar, and bird all give identical *Concept* results for either their singular or plural forms. This is the expected result, given autotruncation.

However, as shown in Table 4, singular versus plural multiword phrases may generate the same, slightly different, or dramatically different results at a *Concept* level.

These results demonstrate the complex interactions of the algorithms that generate the *Concept* results. However, the author can discern no pattern to these results, though they do appear unrelated to whether one or neither form of the phrase happens to match a CAS General Subject Heading, which was an initial hypothesis of the author.

This author highly recommends that CAS modify whatever algorithms are being invoked so that singular and plural forms of chemical synonyms, biological organisms, and simple multiword phrases generate identical results at the *Concept* level. The variance in hit counts between singular and plural forms, regardless of the exact cause, creates confusion for experienced and novice SciFinder Scholar searchers alike. The differences in hit counts demonstrated above are counterintuitive, given that the system generally provides extensive automatic truncation.

VIII. CONCLUSION

Having done hundreds of search variations in preparing this paper, the author has been frequently amazed and at times puzzled by the nuances of the NLQ processing system. Overall, SciFinder Scholar's Explore by Research Topic meets its goal of effectively exploring the scientific literature whether one is a novice or experienced searcher. The search results are nearly always highly relevant and often surprisingly comprehensive regardless of the level of complexity or syntax of the query.

Chemical Abstracts Service's commitment to this interface is demonstrated by new versions at least once every two years and regular adjustments to the NLQ algorithms and associated dictionary/thesaurus equivalents. As noted in section VII, the only major improvement in the NLQ processing system recommended by the author would be to present consistent and identical results for the singular and plural forms of simple queries. SciFinder Scholar continues to be among the best database interfaces available to scientists today.

APPENDIX: SCIFINDER TIMELINE

The following timeline expands information published in a SciFinder database review by Bolek.[18]

1991

• **August** − CAS began to explore the idea of a new desktop research tool.

1992

• **September** − CAS formed a new product development group.

1993

• **July** − The first prototype was created, and α testing at five sites was performed.

1994

• **Summer** − β testing with an additional eight companies was performed.

• **October** − SciFinder (commercial version) was introduced for the first time at Online/CD-ROM '94. It had three simple search options; three research functions; "Explore" for author, chemical substance including structures, and topic searches; and "browse table of contents" and "keep me posted" current awareness.

1995

• **April** − CAS begins abstracting and indexing electronic-only documents, though only a few sources meet CAS's criteria.

• **November** − SciFinder wins the Information Industry Association's HotShots award as "best science/technology service" of the year.

1996

• **March** − CAS announces a task-based pricing option in order to make SciFinder more affordable for small organizations.

• **April** − CAS and Derwent Information announces a strategic agreement to include information from Derwent's World Drug Alerts and Patents Preview via SciFinder.

• **June** − SciFinder version 2 is released with reaction substructure searching, access to supplier information (CHEMCATS), and links to the CAS Document Detective Service.

• **September** − SciFinder wins a "Top 100" most technologically significant new products and processes award from R&D magazine.

1997

• **Fall** − SciFinder version 3 is released, in which a substructure search module was added, and was priced as an add-on to the base product.

• **November** − ChemPort full-text portal debuts at Online '97 with eight participating publishers. It is not yet integrated into SciFinder.

• **December** − SciFinder Scholar version 1 is released. It has four simple search options: chemical substance or reaction, research topic, author name, and document identifier.

1998

• **Fall** − SciFinder version 4 is released including regulatory information (CHEMLIST) and ChemPort linking to full-text articles.

• **September** − The SciFinder Substructure Search Modulate wins a second "Top 100" most technologically significant new products and processes award from R&D magazine.

• **November** − SciFinder Scholar version 2 is released, which links to full-text articles added via ChemPort.

1999

• **January** − CAS begins adding cited references to CAplus records.

• **February** − The SciFinder Scholar substructure search module is released.

• **March** − SciFinder version 5 is released, which includes MEDLINE, a full reaction query tool, and patent family information.

• **September** − SciFinder Scholar version 3 is released, which includes a full reaction query tool.

2000

• **January** − CAS begins covering preprints in CAplus.

• **February** − SciFinder Scholar is enhanced with chemical suppliers (CHEMCATS) and regulated chemicals information (CHEMLIST), which had been available to commercial customers previously.

• **Fall** − SciFinder 2000 is released, which includes company name searching, a Panorama analysis tool, and functional group transformations.

• **Fall** − SciFinder Scholar 2000 is released; CAS began using the year for version numbering and synchronized SciFinder and SciFinder Scholar version numbers. MEDLINE, table of contents browsing, company name searching, and cited references are all added.

2001

• **April** − SciFinder products can now access 1947 to present Chemical Abstracts records.

• **Fall** − SciFinder 2001 is released; biosequence searching, calculated property data, reaction information back to 1974, and links to cited and citing references are added.

• **Fall** − SciFinder Scholar 2001 is released, which includes links to cited and citing references, reaction information back to 1974, and calculated property data.

2002

• **February** − SciFinder products can now access 1907 to present Chemical Abstracts records.

• **August** − SciFinder 2002 is released. It includes stereospecific structure searching, experimental property data, and reactions back to 1907.

• **August** − SciFinder Scholar 2002 is released, which includes stereospecific structure searching and additional analyze and limit options.

• **December** − The addition of subject and substance indexing to records for the seventh Collective Index (1962−1966) is completed.

2003

• **May** − The addition of subject and substance indexing to records for the sixth Collective Index (1957−1961) is completed.

• **August** − The addition of subject and substance indexing to records for the fourth and fifth Collective Indexes (1937−1956) is completed. SciFinder 2004 is released (the traditional fall release is now numbered for the coming year just like automobiles). It includes a new substance alerting feature and several additional options for analyzing and limiting queries and answer sets.

• **October** − The addition of subject and substance indexing to records for the first through third Collective Indexes (1907−1936) is completed.

2004

• **February** − The addition of experimental properties to substances records is begun.

• **October** − Select 1900−1906 literature references are added to CAplus.

2005

• **August** − SciFinder 2006 for Windows is released. It includes structure similarity, variable point of attachment, and repeating group for structure searching. Duplicate citation removal, find specific reference, and main menu redesign are three more new features. Nearly 14 000 records dating from 1879 to 1900 were added to CAplus from the *Journal of the American Chemical Society* and *Journal of Physical Chemistry*, including patents and other journal articles abstracted in these two publications.

• **November** − The native MAC OS X version of SciFinder 2006 is released.

REFERENCES AND NOTES

(1) Somerville, A. N. SciFinder Scholar. *J. Chem. Educ.* **1998**, *75*, 959, 975−976.
(2) Williams, J. SciFinder from CAS; Information at the desktop for scientists. *Online (Medford, NJ, U. S.)* **1995**, *19*, 60−66.
(3) Chemical Abstracts Service. What is New in SciFinder Scholar 2006 [User Guide Home Page]. http://www.cas.org/SCIFINDER/help/2006/SCH_Help/Ctxt_eps/default.htm (accessed Dec. 14, 2005).
(4) Ridley, D. D., *Information Retrieval: SciFinder and SciFinder Scholar*. Wiley: Chichester, West Sussex, U.K., 2002.

(5) Nitsche, C. I.; Buntrock, R. E. SciFinder 2.0: Preserving the partnership between chemist and information professional. *Database* **1996**, *19*, 51−54, 56−58.
(6) Schwall, K.; Zielenbach, K. SciFinder: a new generation of research tool. *Chem. Innovation* **2000**, *30*, 45−50.
(7) Haldeman, M.; Vieira, B.; Winer, F.; Knutsen, L. J. S. Exploration tools for drug discovery and beyond: Applying SciFinder to interdisciplinary research. *Curr. Drug Discovery Technol.* **2005**, *2*, 69−74.
(8) Whitley, K. M. Analysis of SciFinder Scholar and Web of Science citation searches. *J. Am. Soc. Inf. Sci. Technol.* **2002**, *53*, 1210−1215.
(9) Rosenstein, I. J. A literature exercise using SciFinder Scholar for the sophomore-level organic chemistry course. *J. Chem. Educ.* **2005**, *82*, 652−654.
(10) O'Reilly, S. A.; Wilson, A. M.; Howes, B. Chemical information instructor: Utilization of SciFinder Scholar at an undergraduate institution. *J. Chem. Educ.* **2002**, *79*, 524−526.
(11) Eppley, H. Integration of Chemical Information Exercises in an Upper Level Inorganic Lab Course. *Abstracts of Papers*, 36th Central Regional Meeting of the American Chemical Society, Indianapolis, IN, June 2−4, 2004; American Chemical Society: Washington, DC, 2003; GEN 324.
(12) Brannon, K. L.; Schenck, R. J.; Toler, L. S. Reaction information discovery using CAS' SciFinder and SciFinder Scholar. *Abstracts of Papers, Part 1*, 226th National Meeting of the American Chemical Society, New York, Sept 7−11, 2003; American Chemical Society: Washington, DC, 2003; CINF 101.
(13) Ridley, D. D. Strategies for chemical reaction searching in SciFinder. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1077−1084.
(14) Bremner, J. B.; Castle, K.; Griffith, R.; Keller, P. A.; Ridley, D. D. Mining the Chemical Abstracts database with pharmacophore-based queries. *J. Mol. Graphics Modell.* **2002**, *21*, 185−194.
(15) Ridley, D. D. Introduction to structure searching with SciFinder Scholar. *J. Chem. Educ.* **2001**, *78*, 559−560.
(16) Shively, Eric. Chemical Abstracts Service, Columbus, OH. Personal Communication, 2005.
(17) Macko, John. Chemical Abstracts Service, Columbus, OH. Personal Communication, 2006.
(18) Bolek, A. D. SciFinder Scholar [Database Reviews and Reports]. *Issues Sci. Technol. Libr.* [Online] **2000**, No. 28.

CI050481B