Coach™: applying UMLS Knowledge Sources in an expert searcher environment

By Lawrence C. Kingsland III, Ph.D.
Chief, Computer Science Branch

Anna M. Harbourt, M.L.S.
Information Research Specialist, Computer Science Branch

Edmund J. Syed, B.S.E.E.
Electronics Engineer, Computer Science Branch

Peri L. Schuyler, M.L.S.
Head, MeSH Section

National Library of Medicine
8600 Rockville Pike
Bethesda, Maryland 20894

With the development of the Unified Medical Language System™ (UMLS®) Knowledge Sources, the National Library of Medicine (NLM) has produced a resource of great potential for improving the searching of MEDLINE.® The Coach™ expert searcher system, an in-house research project at NLM, is designed to help users of the GRATEFUL MED® front-end software improve MEDLINE search and retrieval capabilities. This paper describes the Coach program, the knowledge sources it uses, and some of the ways it applies elements of the UMLS Metathesaurus® to facilitate access to the biomedical literature.

INTRODUCTION

Since the early 1970s, the National Library of Medicine (NLM) has made searching the world's biomedical literature faster and easier by providing information retrieval on the MEDLARS® family of databases. MEDLINE,® the largest and most frequently used of these databases, comprises over 7,000,000 citations and is searched more than 18,000 times a day on the NLM system alone. NLM's system is estimated to account for approximately 30% of total online use of MEDLINE worldwide. The number of new user accounts and passwords issued by NLM for use of its online system has increased very significantly in recent years. Almost all the new passwords are being issued to new end users [1]. To facilitate end-user access to these millions of citations, NLM developed a PC-based front-end program called GRATEFUL MED.® GRATEFUL MED, which became available in 1986, now has more than 40,000 users. The GRATEFUL MED program and its use have been reported in many articles elsewhere [2–5].

As GRATEFUL MED has evolved to version 6 for the PC and version 1.5 for the Apple Macintosh,® NLM has added significant new functionality in response to users' needs. Now, an adjunct program called Coach™ is being developed by NLM researchers to help the GRATEFUL MED user improve retrieval from MEDLINE [6]. Coach is an expert searcher system with the primary function of applying the Unified Medical Language System™ (UMLS®) and other specialized resources to revise suboptimal searches and improve end-user access to the biomedical literature.

In building the Coach program, NLM has tried to emulate the approach of an expert human searcher in diagnosing search problems and applying specialized knowledge to help resolve them. The user tells the system whether to increase retrieval or to focus and limit retrieval. Coach analyzes the user's search, interacts with the user by applying or suggesting alternative mappings from its knowledge sources, and then invokes GRATEFUL MED to submit the revised search back to NLM's ELHILL® mainframe retrieval system.

Coach knows enough about GRATEFUL MED and the ELHILL command language to create effective searches. The fact that the Metathesaurus, GRATEFUL MED, Medical Subject Headings (MeSH®), MEDLINE, and ELHILL are themselves all moving targets, continually being expanded and improved, has made the development of the Coach system somewhat challenging.

BACKGROUND

Some of the initial analysis which led to this project began more than five years ago. In April 1987, there were 504 MEDLINE searches conducted using GRATEFUL MED, version 2. When these 504 searches were analyzed with the users' permission, NLM found that 37% resulted in null retrieval. Of the null retrievals, more than half—51%—were caused by people "AND"-ing themselves into no return. They entered the logical equivalent of "A AND B, AND C, AND D . . . ," and the result was zero retrieval because the intersection of all these sets was empty. Of the null retrievals, 49% were "no postings" searches: a search statement simply found nothing. Of the "no postings" searches, 30% were problems with author searches. Spelling problems accounted for 25% of the "no postings" searches; 22% were attributable to punctuation or truncation errors; and 12% were failed title searches. Clearly, there was room for improvement. Walker [7] and Mitchell [8] have examined more recently some of the areas in which users have difficulty searching with GRATEFUL MED.

When looking at current information on the use of GRATEFUL MED, substantial changes have taken place. GRATEFUL MED has grown and improved since 1987, and users are better informed. In September 1992, there were 245,521 search sessions by users of all versions of GRATEFUL MED, including Macintosh versions 1 and 1.5 and PC versions 1–6, an impressive 487-fold increase in five-and-a-half years. Of the 245,521 searches, 27% retrieved no citations. Although this represents a significant improvement over the original 37%, it still could be better.

Fifty-seven percent of the failed GRATEFUL MED searches conducted in September 1992 were caused by users "AND"-ing terms for zero retrieval. The search statements had postings, but the intersection was null. This represents an increase from 51% in the searches conducted in 1987. Of the failed searches in September 1992, 23% are "no postings" searches, a significant improvement over the 49% in 1987. The most likely explanation for this improvement is that the MeSH vocabulary used for indexing MEDLINE is only an F10 key away in GRATEFUL MED. GRATEFUL MED users actually do seem to be using the F10 key and searching with MeSH headings [9].

In the final analysis, it could be argued that there are really three categories of problems. The users retrieved too little or none at all; they retrieved too much—a lesser problem, but real if they wanted 14 citations and got 3,012; or they retrieved inappropriate citations. There are many ways users can be helped by a program which knows some of the common ways searches can fail and has knowledge sources to help the user map to better terms [10–11]. The most important single contribution of a program like Coach (potentially improving 57% of all the failed searches) would be to address successfully the problem of Boolean combinations which result in null retrieval.

COACH KNOWLEDGE SOURCES

Coach is one of the first PC-based programs to use the UMLS Metathesaurus® to augment user search terms and help find new terms. It can invoke as many as eleven different knowledge sources to assist in diagnosing problems, mapping to new terms, and otherwise attempting to improve suboptimal searches. The current test version of Coach, version 1.0, uses the Meta-1.1 release of the UMLS Metathesaurus. Meta-1.2 will be incorporated as soon as it becomes available. From the Metathesaurus, Coach knows which concepts came from MeSH. For each of these concepts, Coach knows which terms are explodable and which have pre-explodes available in ELHILL. It knows the topical subheadings and the subheading qualifiers allowable with each Metathesaurus concept which is a MeSH term. It also knows the MeSH "consider also" terms, the "see related" forward cross-references, the main heading/subheading combinations, and the check tags. Several examples of mappings from these and other knowledge sources used by Coach are presented as an appendix to this paper.

Coach has other knowledge sources which map synonyms of subheadings to subheadings and pre-exploded subheading clusters; provide professional specialty headings to search terms more likely to represent the user's intended query; track GRATEFUL MED's GMTERMS.SYN file so Coach does not repeat actions GRATEFUL MED will itself perform; and identify query terms which are ELHILL stop words. For users unsure of the exact spelling of terms, Coach can offer "Soundex" spelling help.

COACH-ASSISTED SEARCHES

Examples of assisted search revisions by the Coach program are presented. Several deal with the biggest single problem in failed GRATEFUL MED searches: users "AND"-ing themselves into null retrieval. The Coach development team has concentrated on the problem of getting useful retrieval with a search that initially returned nothing from MEDLINE.

Assisted increase

Coach has an "assisted increase" mode in which it works through a series of ten diagnostic steps and remedies to improve a failed search. The user can follow the process with a built-in "road map" to show what step is currently active. It can be enlightening and quite interesting to turn on the road map and watch the program hunting for ways to fix searching problems.

In the first sample search, the user is interested in articles about stress fractures of the spine, uses the MeSH F10 key in GRATEFUL MED, and finds "FRACTURES, STRESS" and "SPINE." The logical "AND" of these perfectly reasonable-sounding terms in MEDLINE retrieves only one hit and a very surprised user. So the user says, "Time out; call the Coach." Coach is invoked from GRATEFUL MED version 6 with one keystroke. It analyzes the user's original search and the response from ELHILL. Because there was only one hit, Coach assumes the user probably wants more and pre-positions its command cursor on "assisted increase." The user who does want more presses the "enter" key, and Coach begins the ten-step process of trying to increase retrieval. It discovers that the search term "SPINE" has seven narrower terms in the MeSH hierarchy and includes them in the search with the "explode" command. Coach looks further and finds that the term "SPINE" maps to the "consider also" knowledge source. Coach offers to include those terms which start with "vertebral:" and "spondyl:." The user selects both, and Coach adds them to the search. The program offers to resubmit the search at that point, but the user wants Coach to try harder. Coach passes "FRACTURES, STRESS" to the Metathesaurus browser, and that concept heads the browser's retrieval list. Number two on the list is "FRACTURES, SPONTANEOUS," which the user likes and selects. Coach adds it to the search, and the process is complete. The resulting search retrieves seventy-eight hits in MEDLINE—much better than the one hit of the original search.

Another example of an "assisted increase" uses Coach to help a user interested in the effect of Lyme disease on vision. A GRATEFUL MED search on "Lyme disease" and "vision" gets no hits at all. The problem is not that MEDLINE has no citations on this topic, but that the user is not querying the system properly.

Coach finds that both *Lyme disease* and *vision* have narrower terms in the MeSH vocabulary, so it explodes both of them. It spots the term "VISION," and realizes that that term has a "see related" forward cross-reference to "VISUAL PERCEPTION." It offers to augment the search with "VISUAL PERCEPTION," and the user says yes. With the user's permission, Coach runs the revised search and retrieves eleven relevant hits.

Assisted focus

Coach also has an "assisted focus" function with multiple steps to help the user narrow a search to fewer but better hits. As with other Coach functions, some parts of the process are explicitly interactive: the user chooses from among the options. Some parts are fully automatic: the expert human searchers interviewed in the knowledge engineering process told the Coach team, "Just do it."

The user has searched on "AZT" and "AIDS," wanting to know about the use of Zidovudine in the treatment of acquired immunodeficiency syndrome. The first try yields 808 hits—more than the user wanted. The user wants to focus the search, narrowing retrieval with the "assisted focus" capability. Coach helps the user add the "central concept" restriction to limit retrieval to articles in which the terms were judged by the indexers to be a key concept. Then Coach offers explicit guidance in adding subheadings as qualifiers to each of the search terms. The usage of each subheading, allowed as a term qualifier, is explained on the screen as the user moves down the scrollable subheading "pick list." The user adds the central concept restriction and the subheading "THERAPEUTIC USE" to "AZT." The user adds the central concept restriction and the subheading "DRUG THERAPY" to "AIDS." The resulting search retrieves 186 hits in MEDLINE. The user, still wanting fewer hits, returns to Coach to add "limit" options. Coach offers pick list access to seven limit options: language, publication type, publication set, search years, current month (SDILINE®), check tag, and age group. The user chooses language and publication type and limits the search to English and publication type to articles describing clinical trials. The result is thirty-one very good, tightly focused hits.

COACH METATHESAURUS BROWSER

The Coach Metathesaurus browser is the key that provides users access to this resource. The Metathesaurus, a product of the UMLS initiative, contains hundreds of thousands of terms. Definitions, lexical variants, synonyms, related terms, co-occurrences of terms with other terms in articles indexed in MEDLINE, semantic type assignments, previous indexing for MeSH-derived terms, "broader than/narrower than" relationships, and many other elements are present. Through the Metathesaurus, Coach can map the user's term to related terms in MeSH and other vocabularies.

The Coach browser is a non-Boolean retrieval engine that accepts multiple term input, runs against the entire Metathesaurus, and produces a ranked list of Metathesaurus concepts as output. The browser presents these concepts in the form of a scrollable

pick list. The user can select terms and bring them back to augment a search. The Metathesaurus concept definitions are presented with the concept pick list on the screen and change to follow as the user moves down the list. Tree contexts, single or multiple, are on the screen for those concepts which came from MeSH, *CPT* (*Current Procedural Terminology*), or *ICD-9-CM* (*International Classification of Diseases, 9th edition, Clinical Modification*), all of which have a hierarchical structure.

In Coach's Metathesaurus browser, the user is one keystroke away from the narrower terms of the highlighted Metathesaurus concept, and another keystroke away from the sibling terms at the same level of indentation in the tree hierarchy. Both are also presented as pick lists from which the user can select individual terms and bring them back to the search. The UMLS semantic type or types assigned each concept are directly accessible. When the source of the concept was MeSH, the terms under which that concept was previously indexed are available.

Another browser display screen shows Metathesaurus fields containing the concept's synonyms, lexical variants, reviewed related terms, and unreviewed related terms in vertically scrolling fields. Some of these fields have dozens or even hundreds of entries.

The Coach browser's retrieval engine operates from a universe of 92,461 main concepts or synonyms, 78,464 lexical variants, 32,090 previously indexed terms, 33,347 reviewed related terms, and 142,641 unreviewed related terms. The total number of Metathesaurus entries against which the browser operates is 379,003. Since many of these entries are multiword terms like "ERYTHEMA CHRONICUM MIGRANS," the total number of words involved is just over half a million.

The browser's retrieval engine finds hits in the Metathesaurus and ranks their appropriateness. The Coach ranking algorithm weights Metathesaurus record, lexical variant, synonym, previously indexed, reviewed related, and unreviewed related terms. Versions which also used "broader than/narrower than" factors in weighting have been implemented. Inverse term frequency enters into the calculation to help avoid frequent overweighting and the appearance of nonspecific terms.

The Metathesaurus browser has a "why this hit?" function that can be invoked to show in clear matrix form which of the ranking algorithm weighting factors hit for each word of a multiword argument like "ISLETS OF LANGERHANS." This can be helpful with a non-Boolean system, when a puzzled user wonders why something unusual looking appeared in the hit list.

Careful attention to data structures and serious thought on precomputed indexes to gain speed have resulted in a standard response time of one to two seconds for the Coach browser running on a 25-MHz 80386 microcomputer against the entire Metathesaurus with a two-word query such as "mental retardation." That query returns one hundred Metathesaurus concepts as hits; the elapsed time includes ranking. Most queries are faster. The Coach program has logic which tells it how to incorporate terms the user has brought back from the Metathesaurus into the search being revised—when to "OR" the new term and when to "AND" it, depending on what the user was doing at the time.

In the near future, work with some of the more unusual attributes of the Metathesaurus will begin. Coach will make active use of the semantic type assignments, of the allowable relationships between semantic types, and of the Metathesaurus co-occurrence data. When it does, the breadth of this UMLS resource will facilitate, in even more substantial ways, user searching in MEDLINE and other complex databases.

DISCUSSION

NLM's Coach expert searcher system has been described briefly. Its capabilities for providing access to the rich diversity of the UMLS Metathesaurus have been discussed. Examples of the Coach program revising failed MEDLINE searches have been presented. The system offers other functions to save, restore, view, and modify existing searches. The user can search at several stages of the assisted revision process or can prompt Coach to "try harder" until it has done all it knows how to do.

Using the GRATEFUL MED Standalone Search Engine®, Coach can, as a last resort, try different leave-one-out combinations of three- and four-term Boolean "AND" searches which retrieve nothing, in one log-in, without downloading results. The user views the results, makes a choice, and with one keystroke executes the chosen search with download. When appropriate, Coach does true ELHILL multifile searches in MEDLINE and its backfiles. It offers access to the ELHILL SUPERPRINT sorting capabilities, allowing such things as sorts by author and title or by journal, in ascending or descending order. It gives the user substantial flexibility in specifying the format of the download, with online messages to explain commands such as "print," "print full," and "print detailed." Users can also use the "select own choice" function in Coach to pick any combination of fourteen printable elements in a MEDLINE record, including Chemical Abstracts Service (CAS) registry number, gene symbol, and secondary source.

The current version of Coach, working with the Meta-1.1 release of the Metathesaurus, has forty-eight megabytes of local files for its knowledge sources. The present system is written in Microsoft C and runs

on MS-DOS machines. For work group access during testing, without requiring substantial local disk resources at individual PCs, NLM is delivering Coach in-house from Novell file servers.

FUTURE PLANS

The Coach program is now at version 1, entering alpha testing at NLM. It will go to beta test outside NLM late in 1992. Comments and feedback by research collaborators during the testing process will help the development team refine the system's capabilities and add enhancements requested by users. Suggestions from colleagues within and outside NLM will guide the developers in improving the functionality of the Metathesaurus browser, both with Coach and for stand-alone access to this diverse storehouse of knowledge. When the Coach system is helping GRATEFUL MED users actively search MEDLINE more effectively, the Coach group will turn to the implementation of expert searching capabilities for other files and other search areas, such as TOXLINE®.

ACKNOWLEDGMENTS

Most substantial projects are the result of the creative ingenuity and hard work of a number of contributors. The Coach system is no exception. The idea and impetus for the program originated with NLM's director, Dr. Donald Lindberg. An NLM-wide group of colleagues, chaired by Betsy Humphreys, made the requirements analysis which led to the prioritized list of functions Coach should address. Elizabeth Banzer, then an NLM library associate, contributed materially to the process by which Coach analyzes and improves failed searches. Annette Nahin and Catherine Soehner of the MEDLARS management section were very helpful in defining and discussing search problems commonly encountered by GRATEFUL MED users. Dr. Tamas Doszkocs participated in the early system design, particularly of the Metathesaurus browser module. Many NLM staff members have helped by testing and critiquing the system at various stages in its development. To them all, we owe thanks.

REFERENCES

1. Wallingford KT, Selinger NE, Humphreys BL, Siegel ER. Survey of individual users of MEDLINE on the NLM system. Springfield, VA: National Technical Information Service, 1988.
2. Haynes RB, McKibbon KA, Walker CJ, Ryan NC et al. Online access to MEDLINE in clinical settings. A study of use and usefulness. Ann Intern Med 1990;112:78–84.
3. Proud VK, Schmidt FG, Johnson ED, Mitchell JA. Teaching human genetics in biochemistry by computer literature searching. Am J Hum Genet 1989;44:597–604.
4. Mitchell JA, Proud VK, Johnson ED. Teaching medical students to search MEDLINE for genetics and biochemistry. In: Salamon R, Protti D, Moehr J, eds. Proceedings: Medical Informatics and Education International Symposium, May 15–19, 1989, Victoria, British Columbia, Canada. Victoria: Queen's Printer, 1989:202–5.
5. Burroughs CM. Clinicians' satisfaction with GRATEFUL MED: an exploratory study. Bull Med Libr Assoc 1989;77(1): 56–60.
6. Kingsland LC III, Syed EJ, Lindberg DAB. Coach: an expert searcher program to assist GRATEFUL MED users searching MEDLINE. In: Lun KC, Degoulet P, Piemme TE, Rienhoff O, eds. MEDINFO 92: proceedings of the Seventh World Congress on Medical Informatics, September 6–10, 1992, Geneva, Switzerland. Amsterdam: North-Holland, 1992:382–6.
7. Walker CJ, McKibbon KA, Haynes RB, Ramsden MF. Problems encountered by clinical end users of MEDLINE and GRATEFUL MED. Bull Med Libr Assoc 1991;79(1).
8. Mitchell JA, Johnson ED, Hewett JE, Proud VK. Medical students using GRATEFUL MED: analysis of failed searches and a six-month follow-up study. Comput Biomed Res 1992;25(1):43–55.
9. Mitchell JA, Johnson ED, Proud VK. New thoughts about medical students as effective searchers of MEDLINE. Acad Med 1990;65(7):434–7.
10. Sewell W, Teitelbaum S. Observations of end-user online searching behavior over eleven years. J Am Soc Inf Sci 1986;37(4):234–45.
11. Slingluff D, Lev Y, Eisan A. An end-user search service in an academic health sciences library. Med Ref Serv Q 1985;4(1):11–21.

Received October 1992; accepted November 1992

APPENDIX

The following are examples from the Professional Specialty Headings knowledge source, referring the user to other terms, rather than the professional specialty heading the user probably did not mean:

User's term is 'Adolescent Psychiatry.' Mapping ORs in 'explode Adolescence AND the pre-explode of Mental Disorders.'

User's term is 'Audiology.' Mapping ORs in 'explode Audiometry OR explode Hearing Tests.'

Examples from the Synonyms of Subheadings knowledge source, spotting terms the user has entered that are better searched in other ways:

User's term is 'Management.' Offer to substitute either the subheading 'Organization and administration' or the subheading 'Therapy.' If the user chooses 'Therapy,' enter the pre-exploded therapy cluster of subheadings.

User's term is 'Anatom:,' 'Morphol:' or 'Histol:.' Offer to substitute the subheading 'Anatomy & Histology.'

User's term is 'Abnormality,' 'Abnormalities,' 'Agenesis,' 'Anomaly,' 'Anomalies,' 'Defect,' 'Defects,' 'Deformity,' 'Deformities,' 'Malform:,' 'Teratogen:,' or 'Teratol:.' Offer to substitute the subheading 'Abnormalities.'

Examples from the Main Heading/Subheading Combination knowledge source, referring the user away from an invalid MeSH heading and subheading combination to the preferred precoordinated heading expressing the equivalent concept:

Do not use 'Accidents' with subheading qualifier 'prevention & control.' Instead, use the new term 'Accident Prevention.'

Do not use 'Aorta' with the subheading 'radiography.' Instead, use the term 'Aortography.'

Examples from the Consider Also knowledge source, referring the user to other headings which relate to the topic linguistically:

Topic is 'Brain.' Consider Also terms at 'cerebr:' and 'encephal:.'

Topic is 'Bone Marrow.' Consider Also terms at 'myel:.'

Examples from the See Related forward cross-reference knowledge source, referring the user to other headings that relate to the topic conceptually:

Topic is 'Naval Medicine.' See Related 'diving.'

Topic is 'Allergens.' See Related 'bites and stings'; 'dust'; 'feathers.'