Automated MeSH Indexing of the World–Wide Web

Jerry Fowler, Ph.D.[†]
Srinivas Maram[‡]

Vram Kouramajian, Ph.D.[‡*]
Vijayanth Devadhar[‡]

[†]Department of Community Medicine
Baylor College of Medicine
Houston, Texas 77030
gfowler@bcm.tmc.edu

[‡]Department of Computer Science
Wichita State University
Wichita, Kansas 67260
{*vram,sxmaram,vsdevadh*}*@cs.twsu.edu*

To facilitate networked discovery and information retrieval in the biomedical domain, we have designed a system for automatic assignment of Medical Subject Headings to documents retrieved from the World–Wide Web. Our prototype implementations show significant promise. We describe our methods and discuss the further development of a completely automated indexing tool called the "Web–MeSH Medibot."

INTRODUCTION

Retrieval and management of information across computer networks is increasingly important to biomedical researchers, educators, and clinicians. As medical research expands human knowledge in size and complexity, the task of staying abreast of developments in one's field consumes more and more time and resources.

We have designed an automated agent that we call the Web–MeSH Medibot, whose function is to index biomedical pages found on the World–Wide Web (or simply, "the Web") using the Medical Subject Headings (MeSH) of MEDLINE as derived from the Unified Medical Language System (UMLS) Metathesaurus. This Medibot employs several diverse networked medical resources to build a medical knowledgebase and use it to annotate biomedical Web pages. This paper describes the problem of navigating networked hypertext and the approach used by our Medibot to assist in information discovery, as well as preliminary results of testing a prototype. We discuss related work in categorization of both medical literature and the Web, and conclude with some remarks about integrating this tool with other categorization and information discovery tools to support biomedical researchers in their efforts to keep abreast of the rapidly burgeoning volume of electronically published literature.

Information Retrieval and the Internet

Collaborative retrieval and management of medical information by computer network is an ongoing challenge in medical informatics; however, the vast majority of biomedical researchers and clinicians rightfully see computers not as a central research topic, but as mere tools to expedite their own work. Regrettably, the use of computers often demands substantial "cognitive overhead," that is to say, learning and using the system itself may require so much attention to the details of its operation that the time devoted to the original purpose, the accomplishment of some intellectual task, can be diminished. To create a networked computing environment that improves the productivity of scholars and healers is a major challenge.

A valuable tool for enabling non-computerist users to collaborate is *hypertext*, which is computer-supported document manipulation in which links between one document or piece of a document and another can be traversed automatically, relieving the user of the burden of manual lookup of definitions, footnotes, references, or other related material. Hypertext can be called a non-computerist's interface for data engineering and information management. In contrast to traditional databases, hypertext allows users to manage information informally, developing structure as they learn more about their data. This reduces the cognitive overhead of their information management, allowing them to focus on the information itself.

The rapid increase in use of the Web since its inception in 1991 is a clear demonstration of the utility of distributed hypertext in scientific research and collaboration [1]. Using a system originally developed at the European high–energy physics laboratory, CERN, the Web has grown exponentially in size since the release in 1992 of the first popular graphic user interfaces for Hypertext Markup Language (HTML, the formatting language for Web documents). Personal research libraries, departmental publications, and information retrieval engines are all appearing on the Web. Today's explosion of individually produced publications dis-

[*]Current Affiliation: Knight-Ridder Information, Inc., 2440 El Camino Real, Mountain View, CA 94040.

0195–4210/95/$5.00 © 1995 AMIA, Inc.

tributed via the Web demonstrates that simple, distributed hypertext can engender collaboration across continents and oceans. Among a list of useful Web sites too numerous to mention, the Web server of the National Library of Medicine (NLM), "HyperDOC," is a good example of a significant resource with the potential to make large amounts of knowledge available to many researchers [2].

The UMLS Metathesaurus is a natural hypertext domain and an important resource for medical vocabulary [3]. It contains semantic links that can support the development of semantic inference engines for use in bibliographic as well as clinical and research data retrieval. Grouping objects by semantic locality (that is, categorizing or differentiating them) is a fundamental intellectual activity addressed by the UMLS Metathesaurus, which subsumes the MeSH. Nelson examines several dimensions of semantic locality in the UMLS [3]. Our work concerns itself with conceptual locality (synonymy) and occurrence locality (the co-occurrence of terms in MEDLINE citations and Web pages). We apply these measures to MeSH indexing pages of the Web.

WEB INDICES AND MESH

Exploration of the Internet to obtain information, even using hypertext systems like the Web, has inherent difficulties. Users often experience disorientation as they browse deep into networks of links they are unfamiliar with, a phenomenon commonly referred to as "being lost in hyperspace." The vastness of the Web, with its seemingly countless links, leaves many users without clear notions of what is available or how to find it. Discovery of networked information is the major challenge to using the Web in collaborative research. Among the search engines that first created indices of the Web is the World-Wide Web Worm [4]. More recently, the Harvest system has addressed problems of scalability of search and redundancy in Web indices [5].

An important aspect of search is the ability to assign categories to the texts of bibliographic or other references. Text categorization can be characterized as the exploitation of some desired measure of semantic locality to divide documents into groups. The MeSH addresses this question in medical literature by assigning to each article indexed in MEDLINE a set of terms from a standard vocabulary.

Significant work in automated MeSH-term assignment has been done by Chute and Yang. The general technique involves acquiring a mapping from the vocabulary of each document in set of documents (the training set) to the set of MeSH terms that have been assigned to the document(s) by human experts. This mapping is then used automatically to assign MeSH terms to other documents [6]. Statistically, these methods seem not only to compare favorably with other categorization techniques, but also to offer potential for providing assistance to human indexers [7]. Chute and Yang have used several variations on this general technique for assigning MeSH terms to clinical patient assessments and to automated indexing of MEDLINE citations. The next section discusses the application of these and similar techniques to the development of an indexing system for documents on the Web.

MeSH–Indexing the Web

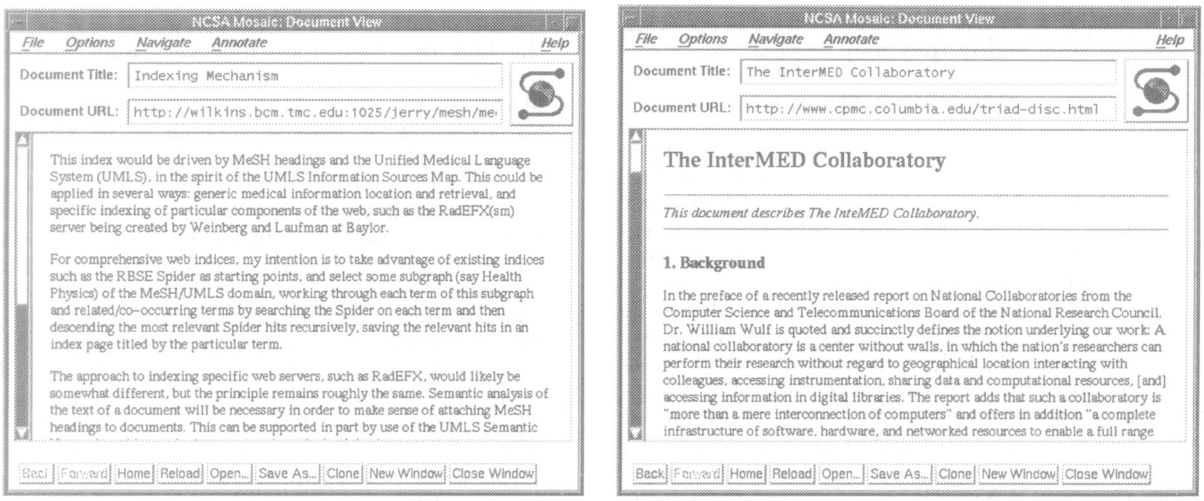
Described briefly, the method used in this paper is to retrieve exemplar citations from MEDLINE for a MeSH term or small set of related terms, build a term-map by applying one or another indexing engine in the manner described by Chute and Yang [6], and then apply that term-map to the text of a Web-page to assign MeSH terms to that page. Two kinds of results can be developed: Web-pages can be retrieved by recursively following the embedded links, and the resulting term assignments can be stored to create a static index of the subset of the Web thus traversed, or pages can be analyzed to attribute MeSH terms to them dynamically, returning only those pages that correspond to a desired term.

The size of the MeSH suggests that it would be prohibitive to provide a term-map for the entire MeSH. For this reason, the development of techniques that use MeSH subsets is appropriate. For the term-map for our tests, we chose the several hundred current MEDLINE citations that were assigned the MeSH term "Information Storage and Retrieval." With these documents, a term mapping technique can be applied to build a term candidate knowledgebase. A Web page can then be indexed using the following term-assignment algorithm.

Term Assignment. To assign indexing terms to a given target page, we retrieve a "similarity set" of those MEDLINE exemplars from our knowledgebase that are determined by our term mapping to resemble most closely the text of the target page. We then retrieve all the indexing terms assigned to each of these papers. This yields the *term candidate set*. A term appears in the term candidate set once for each similar exemplar in which it appears. We then analyze the terms in the term candidate set using a parameterized term-assignment function to assign scores to each term in the term candidate set. Once the scores have been assigned, those terms scoring highest are assigned as terms to the target page.

The term-assignment function, which is summarized here, is essentially similar to that used in [8], to which the reader is referred for more detail. The term-assignment function scores individual terms according to relevance (as assigned by the term-mapping function), frequency of occurrence in related exemplars, age difference between the target page and the similar exemplar, and explicit term weighting in the knowledgebase that provides the indexing terms. In this experiment using MEDLINE, we distinguish between major terms and minor terms as assigned by MEDLINE.

The term candidate set is processed to determine the weighted relevance of each term and normalized to overcome the variations in number of similar ex-



Document Title: Indexing Mechanism

Document URL: http://wilkins.bcm.tmc.edu:1025/jerry/mesh/me

This index would be driven by MeSH headings and the Unified Medical Language System (UMLS), in the spirit of the UMLS Information Sources Map. This could be applied in several ways: generic medical information location and retrieval, and specific indexing of particular components of the web, such as the RadEFX(sm) server being created by Weinberg and Laufman at Baylor.

For comprehensive web indices, my intention is to take advantage of existing indices such as the RBSE Spider as starting points, and select some subgraph (say Health Physics) of the MeSH/UMLS domain, working through each term of this subgraph and related/co-occurring terms by searching the Spider on each term and then descending the most relevant Spider hits recursively, saving the relevant hits in an index page titled by the particular term.

The approach to indexing specific web servers, such as RadEFX, would likely be somewhat different, but the principle remains roughly the same. Semantic analysis of the text of a document will be necessary in order to make sense of attaching MeSH headings to documents. This can be supported in part by use of the UMLS Semantic

Subject Headings
Evaluation Studies
Unified Medical Language System
Software
Information Storage and Retrieval
Algorithms
Comparative Study
Abstracting and Indexing

A

Document Title: The InterMED Collaboratory

Document URL: http://www.cpmc.columbia.edu/triad-disc.html

The InterMED Collaboratory

This document describes The InteMED Collaboratory.

1. Background

In the preface of a recently released report on National Collaboratories from the Computer Science and Telecommunications Board of the National Research Council. Dr. William Wulf is quoted and succinctly defines the notion underlying our work: A national collaboratory is a center without walls, in which the nation's researchers can perform their research without regard to geographical location interacting with colleagues, accessing instrumentation, sharing data and computational resources, [and] accessing information in digital libraries. The report adds that such a collaboratory is "more than a mere interconnection of computers" and offers in addition "a complete infrastructure of software, hardware, and networked resources to enable a full range

Computer Communication Networks
Database Management Systems
Medical Records Systems Computerized
Delivery of Health Care
Unified Medical Language System
Management Information Systems
Hospital Information Systems
Microcomputers

B

Figure 1: MeSH terms assigned to two Web pages using the SMART system

emplars retrieved. Each term in the term candidate set is scored using the function:

$$S(t) = \sum_{e=1}^{n}(R(t,e) + m * r(t,e))$$

where $R(t,e)$ is the normalized relevance of t as a major term in exemplar e, $r(t,e)$ is the normalized frequency of use of t as minor term in exemplar e, and m is the minor term weighting factor (Although one could include a relevance decay factor due to age [8], we have not done so in this test).

After all term candidates have been scored, candidates scoring more than some major–term threshold K are assigned as major terms of the target page. Those remaining terms that score greater than some minor–term threshold k are assigned as minor terms.

Experiments

We used two different experimental approaches to creating our term maps. Both approaches used the same set of MEDLINE citations. In one approach, we used WAIS indexing [9]. Because of the relative inflexibility of the WAIS software we used, and the wide variation in computed relevance of clearly related exemplars, it was not possible to establish a fixed threshold for similarity of an exemplar to a target. We chose to create a dynamic acceptance threshold for similar citations by adding a collection of threshold documents containing "random" medical text. These boundary documents were created by dividing the the entire definitions file of the UMLS Metathesaurus into roughly 90 pieces of contiguous text. We then deemed to be similar any abstract that ranked as more similar than any of these definitions files, with the argument that if an abstract is not any more similar than "random" text, it must not be similar to the target page.

Our second approach employed the SMART system [10]. In this approach, we divided the relevance factor of the document among terms, such that the summation of weights assigned to all major and minor terms is equal to relevance factor of the document. This has the effect of normalizing for number of terms assigned to exemplars, giving preference to terms that are assigned in small numbers over those that are assigned in large numbers. Although a dynamic acceptance threshold might improve results with the SMART system as well, results were, in general, much superior.

Two examples of term assignment using SMART appear in Figure 1. Example A is a Web page that was written to describe this project. Example B is a Web page describing the InterMED Collaboratory [11]. Each term list was manually edited after return from our SMART knowledgebase to remove a class of secondary terms, such as "Support U S Govt P H S." The automatic attribution of this term (as well as others such as "Human") to an arbitrary piece of text is clearly an inappropriate extrapolation for this software to make. For this reason, we are developing a stop–list of MeSH terms that should not be assigned by our software no matter how highly they may score.

DISCUSSION

For reasons of simplicity of prototype development, we chose to use two publicly available indexing tools, WAIS and the SMART system. Chute and Yang have explored two other methods of mapping between texts and MeSH terms: Least Squares Fit (LSF), in which vectors representing the words in text of the training documents are placed in a matrix that is then reduced by single-value decomposition (SVD) [12]; and ExpNet, which finds "nearest-neighbors" to the target text [13]. ExpNet is much faster to train (SVD is an algorithm of time complexity $O(n^3)$; ExpNet training has linear complexity) at the expense of slower term assignment. Chute and Yang have compared their techniques with SMART, demonstrating somewhat superior performance. Our Web indexing might be improved by employing either ExpNet or LSF for term-mapping. The utility of LSF in this application may depend on the size of the training set needed to achieve adequate results. Since this set appears to be fairly small, SVD may be feasible.

Enhancements

We intend to explore several possible enhancements to our method in the interest of improving the precision of our term assignments.

We currently pass an entire Web page as a query to the term-mapping function. This has two drawbacks: First, the terms "head" and "body" are commonly used in HTML to delineate text structure, and may distort the outcome of term mapping if not removed by preprocessing. Second, we lose some nuance by failing to assign different weights for the title, head, and body of a page.

We might also replace links with the page titles to which they point before applying the term-mapping function. We can extend this by assigning terms for each page that is linked from the target page. Applying Categorization by Reference [8], we would expect these linked pages to share common keywords with the the target page; in this way, we might hope to increase our precision in assigning terms to the target.

Experience has shown that the precision of assignment can be improved by making use of the MeSH hierarchy to determine similarity. A target page whose term candidate set is divided among sibling terms may be best characterized by assigning the parent of those terms to the target. Brute-force analysis, however, would be inefficient. We have not yet formulated our approach to this problem.

THE WEB-MESH MEDIBOT

The whole process of building the knowledgebase and indexing individual pages can be automated, using what we call the "Web-MeSH Medibot." "Medibots" are query engines designed to traverse the Internet applying semantic knowledge acquired from the UMLS and other knowledge resources to the task of information discovery and categorization of electronic literature in the medical domain. Medibots are an application of *Amanuensis* (Latin for *secretary* or *copyist*), which is an implementation of the Virtual Object Model. *Amanuensis* supports naming and typing of data in local and remote information sources and computations, thus facilitating both human and automatic search and retrieval in hypertext, and integrating diverse electronic resources into the same user interface. *Amanuensis* "query engines" are servers that provide interfaces to diverse information resources, such as TexSearch (the local copy of MEDLINE in the Texas Medical Center), the UMLS Metathesaurus, the SNOMED nomenclature, and two distributed hypertext systems, the VNS and World-Wide Web. User interfaces to *Amanuensis* were developed from the VNS and Mosaic.

The Web-MeSH Medibot would make use of networked resources including TexSearch and Meta to build a knowledgebase, which it then would use to categorize pages fetched from the Web. Conceptually, the Web-MeSH Medibot would accept a given MeSH term as input, extracting the subtree related to the term from the UMLS Metathesaurus using Meta; it would then retrieve exemplar citations from MEDLINE for each term in the subtree, build a term-map using their abstracts, and then apply that term-map to a set of Web-pages whose Universal Resource Locators were given as input.

In practice, each of the steps required of this Medibot has been performed, but the necessary procedural linkages to perform the whole sequence automatically have not been made. When it is ready, the tool's accessibility to the research and clinical communities will depend on the development of a comprehensible but flexible user interface. The most straightforward way of accomplishing this is simply to embed the Medibot in the "common gateway interface" of a Web server, using HTML forms as query interfaces, and producing the output as HTML to be displayed by the user's personal favorite Web browser. Providing the flexibility to customize the output based on a user's own preferences without presenting excessive complexity will be a significant challenge.

Application

There are two ways to apply the Web-MeSH Medibot: Static indices of certain Webs can be created to assist in organization of information, for instance the online medical syllabi of Baylor College of Medicine's Medical Informatics Education Center. The Medibot can also be used dynamically to explore and categorize portions of the Web, returning the results dynamically as a single virtual page; this will be useful to scientists exploring newly discovered portions of the Web. We intend to experiment with both of these approaches. Our tool can be extended to build links to existing MEDLINE abstracts dynamically, using, for instance, the TexSearch engine of *Amanuensis*.

Several issues will arise in application of the Web–MeSH Medibot:

Depth of search. The Web is effectively inexhaustible, and attempts to traverse the whole web are both doomed to failure and deemed antisocial. Allowing deep searches also involves decisions about what to do when other indices are encountered in the search.

Definition of the term "document". It may be appropriate to consider several interlinked web pages as a single document, even though they are retrieved in separate requests. An abstract approach to categorizing hypertext is "clustering" nodes that are highly related. A "natural cluster" is a cluster whose relationship is based on the number of independent paths between nodes, rather than on their lexical or semantic content [14]. The use of natural clusters (as well as host or domain boundaries) may be appropriate in the context of deciding how big a "document" is.

Intersection with other MeSH subtrees. To use the Web–MeSH Medibot most effectively, it would be useful to provide methods for applying other MeSH terms to the documents analyzed during a search.

CONCLUSION

Our technique of MeSH indexing for the Web shows promise as a tool for discovery and retrieval of biomedical information on the Web, but there remain many issues to be resolved before it can be useful to the medical community at large. We shall continue to explore the possibilities of the Web–MeSH Medibot by building static and dynamic indices of "weblets" at Baylor College of Medicine, including the RadEFX(SM) service of the Center for Cancer Control Research [15] and in educational applications in medical informatics. However, our approach should not be viewed as limited to this activity; it can be applied to any electronic text source. Furthermore, our use of the MeSH as a categorical vocabulary relates to our interest in supporting biomedical research; any well–organized index method for which there exist abundant electronic texts to serve as exemplars of the index terms can be employed in a similar manner.

Success of the Web–MeSH Medibot at Baylor College of Medicine could lead to its use in indexing other sites.

REFERENCES

[1] Berners-Lee T., Cailliau R., Groff J.-F., Pollermann B. World-wide web: The information universe. *Electronic Networking: Research, Applications and Policy*, 51(2):52–58, Spring 1992.

[2] Rodgers R. P. C., et al. Hyperdoc: The national library of medicine (nlm), 1994. URL: `http://www.nlm.nih.gov`.

[3] Nelson S. J., Tuttle M. S., Cole W. G., et al. From meaning to term: Semantic locality in the UMLS metathesaurus. In *Symposium on Computer Applications in Medical Care*, pages 209–213, 1991.

[4] McBryan O. A. GENVL and WWWW: Tools for taming the web. In *Proceedings of the 1st International World Wide Web Conference*, Geneva, Switzerland, May 1994.

[5] Bowman C. M., Danzig P. B., Hardy D. R., Manber U., Schwartz M. F. Harvest: A scalable, customizable discovery and access system. Technical Report CU–CS–732–94, University of Colorado, Boulder, CO, 1994.

[6] Yang Y., Chute C. G. An example–based mapping method for text categorization and retrieval. *ACM Transactions on Information Systems*, 12(2), Apr. 1994.

[7] Chute C. G., Yang Y., Buntrock J. An evaluation of computer assisted clinical classification algorithms. In *Symposium on Computer Applications in Medical Care*, pages 162–166, 1994.

[8] Kouramajian V., Fowler J., Devadhar V., Maram S. Categorization by reference: A novel approach to automated mesh indexing. In *Symposium on Computer Applications in Medical Care*, Apr. 1995. In press.

[9] Schwartz M. F., Emtage A., Kahle B., Neuman B. C. A comparison of internet resource discovery approaches. *USENIX Computing Systems, The Journal of the USENIX Association*, 5(4):461–493, Nov. 1992.

[10] Salton G., Lesk M. E., Harman D., Williamson R. E., Fox E. A., Buckley C. The smart project in automatic document retrieval. In *Proceedings of the 14th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 356–358, 1991.

[11] The InterMED collaboratory. Available as `http://www.cpmc.columbia.edu/\ triad-disc.html`, 1994.

[12] Chute C. G., Yang Y., Evans D. A. Latent semantic indexing of medical diagnoses using UMLS semantic structures. In *Symposium on Computer Applications in Medical Care*, pages 185–189, 1991.

[13] Yang Y., Chute C. G. An application of expert network to clinical classification and MEDLINE indexing. In *Symposium on Computer Applications in Medical Care*, pages 157–161, 1994.

[14] Botafogo R. A. Cluster analysis for hypertext systems. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 116–125, June 1993.

[15] RadEFX(SM) research resource, 1994. URL: `http://radefx.med.bcm.tmc.edu/default.htm`.