MeSHmap: A Text Mining Tool for MEDLINE

Padmini Srinivasan, Ph.D.
School of Library & Information Science
The University of Iowa, Iowa City, IA 52242

Our research goal is to explore text mining from the metadata included in MEDLINE documents. We present MeSHmap our prototype text mining system that exploits the MeSH indexing accompanying MEDLINE records. MeSHmap supports searches via PubMed followed by user driven exploration of the MeSH terms and subheadings in the retrieved set. The potential of the system goes beyond text retrieval. It may also be used to compare entities of the same type such as pairs of drugs or pairs of procedures etc. In addition there is the potential to generate maps of entities (drugs or diseases etc.) such that the strength of the link between two entities in the map represents their similarity as expressed in the MeSH metadata of the MEDLINE documents. Higher level operators have been proposed to support these comparison and mapping functions. This paper motivates and describes MeSHmap. Future work will include user evaluations of the system.

INTRODUCTION

In a 1999 paper Hearst offers an interesting differentiation between the objectives of text retrieval, text mining and web data mining. In it she emphasizes that the key goal of "mining" whether from well structured databases of numeric data or from text collections is the discovery of new knowledge [1]. Although subjectivity necessarily underlies any assessment regarding the "newness" of some given knowledge, the spirit of her statement is clear. In text mining the emphasis is on extracting knowledge that is at the very least not explicitly present in the source database or text collection that is being mined.

We present a prototype application that mines the metadata of MEDLINE documents to yield high level summaries. These summaries are generated by exploiting the manual indexing available in MEDLINE records. In essence, a MeSHmap derived from the MeSH terms and subheadings offers a high level view of a document subset. The intent behind such summaries is not only to support functions such as text retrieval but also to support text mining through exploration without requiring the user to read the underlying documents. MeSHmap is presented as a text mining application since the information generated in the summaries and by the proposed high level operators are not explicitly contained in any single component document.

MESHMAP SUMMARIES

Given the explosion of information in health care it is very difficult for health care professionals, researchers and educators to keep abreast of literature in their domain [2]. This problem is compounded given the increasingly interdisciplinary nature of research and development. Keeping track of areas that might potentially impact one's own domain is extremely challenging especially since it is generally difficult to predict where these influences are likely to come from.

It is well acknowledged that tools able to filter through and facilitate access to the literature are critical. MeSHmap is consistent with this goal in that it offers summaries of retrieved sets that may be used to guide further document retrieval. However, it also goes beyond text retrieval because by looking at a summary alone a reader may obtain an understanding of the key subareas within the set of documents as well as their relative emphases. Moreover, we are very interested in using these summaries to support exploration for text mining. High level operators to compare entities (such as a group of diseases or drugs) by comparing their underlying document sets and to generate maps are possible. These operators further emphasize the "text mining" potential in this application. For example, an unexpected association between two diseases may trigger research in a new direction.

MEDLINE Record: MeSHmaps are derived from the indexing information in the underlying MEDLINE records. Each indexed MEDLINE record contains several descriptors that have been selected by trained indexers from the MeSH classification scheme. In Figure 1, which shows an abbreviated example of a MEDLINE record, the fields tagged with MH are the MeSH descriptors. The phrases following the "/" symbol represent subheadings. Subheadings qualify the MeSH term and specify the particular aspect of the MeSH concept that is present in the document. It may be observed that some of these are tagged with an "*" which indicates that the corresponding MeSH term subheading combination has a major emphasis in the record. There are more than 19,000 main concepts in MeSH and under 100 subheadings. The NLM MeSH browser specifies

1067—5027/01/$5.00

which subheadings are allowed for which MeSH terms.

UI - 21028513
DP - 2001 Jan
TI - Infliximab-associated reversible cholestatic liver disease
AB - Infliximab, a novel therapy for Crohn disease, has been shown to be both safe and effective. We describe etc..
MH - Adult
MH - Antibodies, Monoclonal/*adverse effects
MH - Antirheumatic Agents/*adverse effects
MH - Case Report
MH - Cholestasis/*chemically induced/diagnosis/therapy
MH - Crohn Disease/*drug therapy
MH - Female
MH - Human
MH - Liver/pathology

Figure 1: An Abbreviated MEDLINE Record

MeSHmap Prototype: MeSHmap, a tool written in Java, is designed to support explorations of MEDLINE metadata in order to compare entities and to identify potentially interesting and novel associations. A major emphasis in the design of our prototype is to enable useful interactivity. There are three major interaction phases. We begin with a *search* phase that is followed by an *exploration* phase that leads to a document *display* phase. The user may move freely between phases, although a search is a necessary starting point.

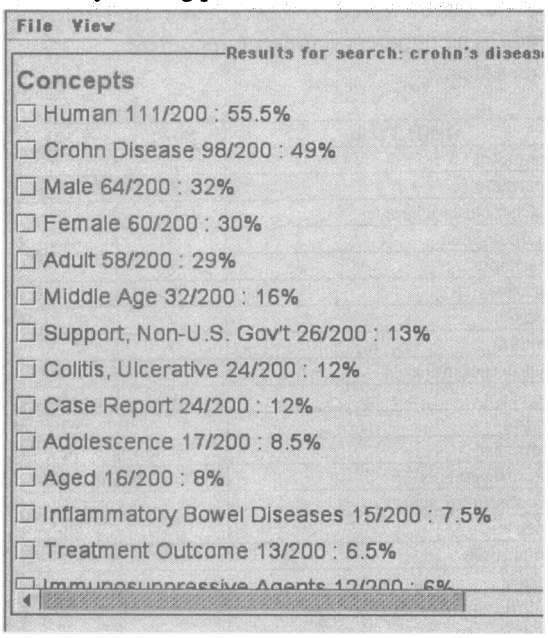

Figure 2: MeSH Concepts in Retrieved Set

Interaction begins by using a search window (not shown) to type in a search criteria. Any search that may be input directly at the PubMed site is valid within MeSHmap. In our example the user is exploring the topic of "Crohns Disease". The system then connects to the PubMed site, hands over the search and downloads the retrieved results. The documents in the result set are analyzed with regards to MeSH terms and subheadings and a summary of this information is provided on the screen. In essence two lists are generated: a list of the MeSH terms (partial screen shot in figure 2) and a list of the subheadings found in the result set (partial screen shot in figure 3). (In order to maintain readability, we show the single window split over figures 2 and 3). For each entry the frequency of occurrence is also provided which allows the user to distinguish sub topics that are core from the others. Thus in our example we may note that out of the 200 documents retrieved, 49% contain "Crohn Disease" as a MeSH term and 10% have some MeSH term qualified by the subheading "adverse effects". At this point the user may explore further details regarding the retrieved set.

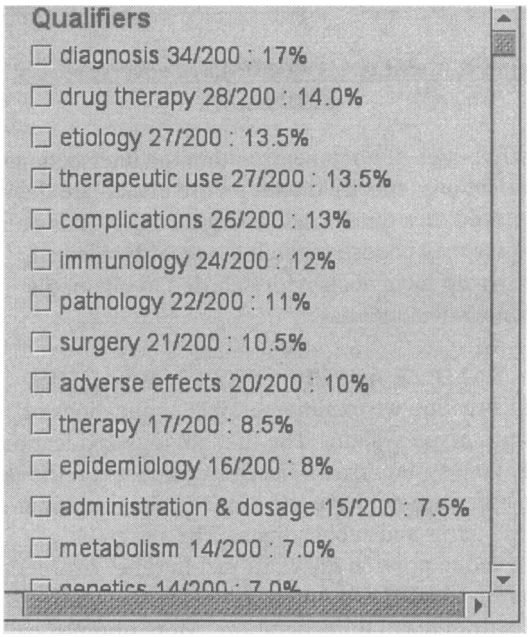

Figure 3: Subheadings (Qualfiers) in Retrieved Set

For example by moving the mouse over a MeSH term, the subheadings associated with that term within the result set and their document frequencies are displayed within a new window. Similarly by moving the mouse over a subheading, the associated MeSH terms are displayed. These operations offer the user opportunities to explore the different sub topics that appear in the retrieved set. It should be noted that all this happens without the user reading the underlying documents. Figure 4 shows the output that is generated when the user moves the mouse over the "adverse effects" sub heading.

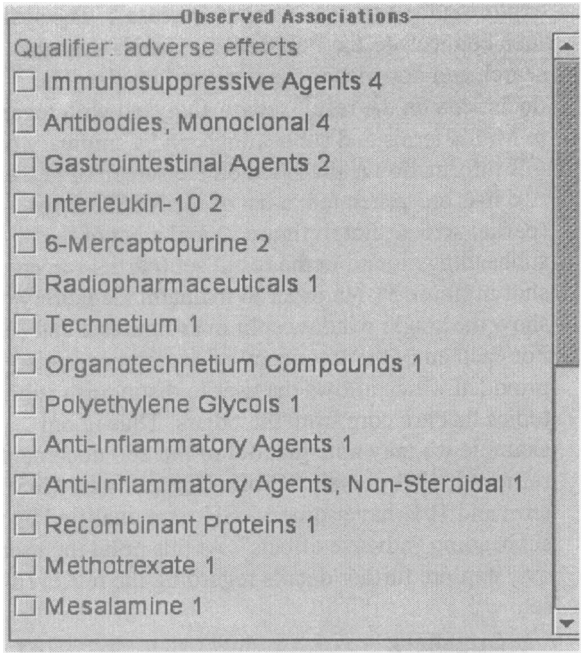

Figure 4: MeSH terms qualified by "adverse effects" in retrieved set.

Finally, titles of documents within the intersections of selections (specified with mouse clicks) are then displayed on a third window (figure 5). At this point the user may choose to fetch the records selected. MeSHmap reconnects with PubMed to obtain the specified document(s).

SAMPLE APPLICATION SCENARIOS

In this section we explore possible applications of this prototype system. The first obvious application is that the system may be used to guide text retrieval. The user does not have to be knowledgeable about MeSH terms and subheadings. The system is designed to provide relevant "just in time" instruction regarding MeSH. At a minimum the onus on the user shifts from recall to recognition. More optimistically, a learning opportunity is offered, wherein the scope of the search topic may be explored via the underlying distribution of its MeSH terms and subheadings.

Scenario 1: Analysis of a Disease over Time: In this scenario, a user may be interested in exploring the progression of ideas in a particular domain, say corresponding to a particular disease. By conducting the search representing the disease twice, each time restricted over different time periods, one may obtain a temporal assessment of the changes in the field. This is done by comparing the MeSH terms and subheadings in the two retrieved sets. There are obvious cautions to observe in such an analysis. Most importantly, indexing practices may have changed across the time periods. Also, we know that the MeSH vocabulary evolves over time. Therefore suitable controls need to be included to raise confidence in conclusions made.

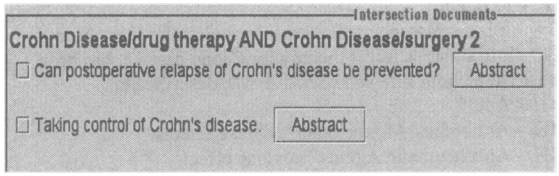

Figure 5: Retrieving Documents from PubMed

We present a preliminary example of this type of analysis. (It should be noted that since this is only an example, controls to accommodate changes in MeSH vocabulary over time have not been included in this longitudinal example scenario). Let us assume that the user wants to explore the evolution of ideas regarding drugs used to treat migraine. The user conducts the search "migraine/drug therapy [MH:NOEXP] AND 1980:1990 [DP] AND clinical trial [PT] AND english [LA]" through MeSHmap. Next the user chooses to explore the MeSH terms in the retrieved set with the subheading "therapeutic use". The same search is repeated but this time the DP value is changed to 1991:2001. The first decade gives 126 documents while the second gives 298 documents. Table 1 shows the 10 most frequent MeSH terms qualified by "therapeutic use" from each decade. The numbers represent document frequencies.

MeSH Term	80-90	91-01
Propranolol	14	
Cinnarizine	12	
Cal. Channel blockers	11	
Acetaminophen	9	
Flunarizine	8	
Piperazines	8	
Naproxen	7	
Clonidine	7	
Dihydroergotamine	7	
Analgesics	7	16
Aspirin		12
Sumatriptan		77
Serotonin agonists		68
Vasoconstrictor agents		36
Indoles		31
Sulfonamides		22
Oxazoles		13
Anti-Inflammatory agents, non-steroidal		12
Analgesics, non-narcotic		11

Table 1: Drug Therapies Explored for Migraine

One can see obvious differences in drug treatments explored in clinical trials across the two decades. For example there is only 1 MeSH term in common between the two lists. Of course, one must also

consider the MeSH hierarchy. However, even the more general categories "Vasoconstrictor Agents", "Indoles", and "Sulfonamides" of Sumatriptan, the highest ranking term in the later decade, do not appear in the top 10 of the previous decade.

In the above example we chose to explore the changes over time of drug treatments studied for a given disease. We may also choose to explore such changes for any other aspect that is supported by the MeSH terms along with the subheadings. For example, one could study changes in the methods used to diagnose diseases by comparing MeSH terms qualified by the subheading "diagnosis".

Scenario 2: Comparison of drugs: Let us assume that at this point the user of scenario 1 decides to compare the drugs Flunarizine and Sumatriptan. A fresh MeSHmap search on "Flunarizine/therapeutic use[MH]" results in a set of 352 documents. Exploring MeSH terms qualified with "drug therapy" within this set will yield almost a total of 110 diseases. That is about 110 different diseases qualified with "drug therapy" co-occur in documents with Flunarizine qualified by "therapeutic use". One may loosely conclude that each of these documents is about the treatment of the corresponding disease with Flunarizine. The same strategy is used to explore Sumatriptan. This time 593 documents are retrieved in response to "Sumatriptan/therapeutic use [MH]". Within this set MeSHmap identifies only 50 unique diseases that are qualified with "drug therapy". Table 2 lists the top 7 diseases for each drug (SU: Sumatriptan and FU: Flunarizine). One may observe that there is an overlap of only 1 disease term in the top 7 lists. Also, there are unique aspects of each drug, for example Flunarizine has been explored in the context of Vertigo, which does not co-occur, with Sumatriptan in any document.

MeSH Term	SU	FL
Migraine	343	68
Headache	45	13
Cluster Headache	40	
Pain	5	
Tension Headache	4	
Myoclonus	4	
Depressive Disorder	4	
Epilepsy		29
Hemiplegia		21
Brain Ischemia		12
Cerebrovascular disorders		10
Vertigo		10

Table 2: Top 7 MeSH Terms Qualified by "drug therapy"

Proposed Comparison operator: The example of scenario 2 motivates the development of a comparison operator that may be used to compare entities of the same type, for example a pair of drugs or a pair of procedures. A query comparing two instances of an entity type is in essence a sequence of two Boolean searches followed by a comparison of their results. In general:Compare(X1,X2,SH1,SH2) =

({Y1},{Y2}, similarity(X1,X2))

where the values for X1 and X2 the two entities being compared and SH1 and SH2 the two subheadings are supplied by the user. Y1 and Y2 are computed as follows:

Y1: {MeSH term t : the search X1/SH1 AND t/SH2 results in at least 1 document}

Y2: {MeSH term t : the search X2/SH1 AND t/SH2 results in at least 1 document} and

similarity(X1,X2) = |{Y1 AND Y2}|/|{Y1 OR Y2}|

Thus the compare operator will return a set of unique MeSH terms for X1 and for X2 that satisfy the above MeSH term/subheading conditions. It also returns the similarity between X1 and X2. It may be noted that in the previous example, the user had specified X1 and X2 to be Flunarizine and Sumatriptan respectively while SH1 was "therapeutic use" and SH2 was "drug therapy".

Proposed Mapping Operator: This operator builds upon the output of the comparison operator. Let us assume that we have a set of entities X = {X1, X2, …., Xn} where the Xis are all of the same type. For example, X could represent a particular family of drugs or a subset of digestive diseases. The goal of Map(X,SH1,SH2) is to generate a map where the nodes represent the entities (members of X) while the links represent their similarities. Thus the link between node Xi and Xj represents the similarity(Xi,Xj). Thus a Map operation begins with a series of Compare(Xi,Xj,SH1,SH2) operations for each pair where Xi is not the same as Xj. The data generated by these comparisons are sufficient to produce the maps.

Strong links in such a map will identify pairs of entities that are well connected in the literature (limited to the constraints of our analysis tools). Displaying the strong connections between the entities has the potential to educate a user who is new to a field. However, the stronger the connection, the less surprising the link may be for a user who is well versed in the subject domain. For such a user, we hope that the weaker connections might actually be more interesting in the sense of exploratory research. If diseases A and B are linked and the domain expert did not expect this, it may provide sufficient motivation for new explorations to understand the reasons. It may be that the user is an expert in only one of those diseases. Since the compare operator will return the sets of terms (Y1 and Y2), it should be possible to display these sets when nodes are selected or the intersections when links are selected.

RELATED RESEARCH

There are numerous examples of research on text mining from MEDLINE. The extraction of protein interactions [3], the interactions between genes [4], genes, drugs and cells [5] etc., define a vibrant research direction. The distinctive aspect about these efforts is that they operate on the text of the MEDLINE record, i.e., the title and abstract. In contrast very few researchers explore mining of the MeSH metadata in MEDLINE. The research of Cimino and colleagues is an interesting exception [6,7,8] . Our work is different in that we propose an exploratory tool, including high level operators, for user-driven text mining from MEDLINE. A related system is ARROWSMITH which has successfully supported the exploration of the subtle, i.e., not explicit, connections between literatures [9]. Our work is similar to ARROWSMITH in that we also focus on text mining across documents.

Exploiting Linked MeSH Terms and Subheadings: Our prototype exploits the implicit relationship between the MeSH term and the subheading. There are risks involved. For instance, since MeSH indexing is done manually, there are no guarantees regarding consistency, completeness or accuracy both within and across documents. However, we know that MeSH terms offer a valuable access method for retrieval (over and above the contributions of the title and abstract fields) [10,11,12]. Thus given their recognized value for text retrieval we suggest that it may also be beneficial to explore their potential in other applications. One option to raise the level of confidence is to limit the analysis to only those MeSH terms and subheadings that are designated as major topics with the asterisk (*).

Hypothesized Relationships Based on Co-occurrence: An assumption may be apparent especially when we discuss the comparison and mapping operators. Essentially we assume the existence of an underlying conceptual relationship tying together a pair of co-occurring MeSH/subheading terms. For instance in example scenario 2 we implicitly assume that if a document is indexed by "Sumatriptan/therapeutic use" and "migraine/drug therapy" then the document discusses the therapeutic use of Sumatriptan for treating migraine. There is the danger of false positive relationships arising from this assumption. Cimino and colleagues have in fact explored this very aspect in several papers [6,7,8]. In [8] they propose a statistical method to distinguish between the significant associations between MeSH terms from the coincidental ones. In future research we plan to incorporate similar methods.

CONCLUSIONS

We present MeSHmap a prototype text mining tool, which operates on the MeSH metadata associated with MEDLINE documents. The prototype supports user driven exploration of MeSH concepts and subheadings in the retrieved set. Our next steps are to implement the comparison and mapping operators. The former may be used to mine relationships between entities while the latter may be used to map sets of entities. In future work we also plan to test the system with users.

Acknowledgements

The author thanks Micah Wedermeyer for implementing MeSHmap and David Eichmann for discussions. This research was supported in part by the NLM grant RO1 LM06909.

References

1 Hearst MA. Untangling Text Data Mining. 1999 Assoc. of Computational Linguistics (ACL) Conference.

2 Gorman PN, Helfand M. Information Seeking in Primary Care: How Physicians Choose Which Clinical Questions to Pursue and which to Leave Unanswered. Medical Decision Making 1995;15(2),113-119.

3 Thomas J, Milward D, Ouzonis C, Pulman S Carroll M. Automatic Extraction of Protein Interactions from Scientific Abstracts.Pacific Symposium on Biocmputing, 2000. 5:538-549, Hawaii.

4 Shatkay H., Edwards S, Wilbur WJ, Boguski,M. Genes, Themes and Microarrays.ISMB, 2000.

5 Rindflesch,TC, Tanabe L., Weinstein JN, Hunter L. EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature. Pacific Symposium on Biocomputing 2000.

6 Cimino JJ, and Barnett GO. Automatic Knowledge Acquisition from MEDLINE. Methods of Information in Medicine, 1993;32(2);120-130.

7 Zeng Q, Cimino JJ. Automated Knowledge Extraction from the UMLS.Proc. AMIA Annual Fall Symp. 1998;568:572.

8 Mendonaa EA, Cimino JJ. Automated Knowledge Extraction from MEDLINE Citations. Proc. AMIA Symposium, 20 Suppl, 2000;575:579.

9 Swanson DR, Smalheiser NR. An Interactive System for Finding Complementary Literatures: a Stimulus to Scientific Discovery. Artificial Intelligence,1997;91;183-203.

10 Srinivasan P. Optimal Indexing Vocabulary for MEDLINE. Info. Proc. and Mgmt. 1996;32(5)}, 503-514.

11 Srinivasan P. Retrieval Feedback in MEDLINE. JAMIA. 1996;3(2)}, 157-167.

12 Aronson AR, Rindflesch TC.Query expansion using the UMLS Metathesaurus. Proc AMIA Annu Fall Symp. 1997;485-9.