Providing Relevant Clinical Information through Agents

Faraz Ahmed and Anwar Usman
SZABIST
Karachi, Pakistan

Abstract: *The size of medical libraries is growing exponentially and data mining agents need to find ways to delve into that information and retrieve the right information which the physician can utilize. This paper discusses a combination of the most effective ways to allow the physician to get the right information in a timely way. It is divided into two parts. Optimizing the input and optimizing the output. The author will combine a few techniques, such as generic queries and query classification, to optimize the input and after the results have been retrieved, will present the output such that the physician has little problem in finding the desired information. Agents are employed to "automate" this task and give each physician the information that he would like.*

Keywords: *Medical information, data mining, agents, medical query optimization.*

1. INTRODUCTION

When searching for information, a user requires the search to deliver results having the following attributes:

i. High Sensitivity: Relevant studies detected out of the total relevant studies.

ii. High Specificity: Non-relevant studies not detected out of the total non-relevant ones.

iii. High Precision: Studies meeting criteria out of the total detected.

iv. High Accuracy: All relevant, detected studies plus all non-relevant, not detected studies divided by the total studies.

While this is the ideal scenario where the user can get all the correct and relevant information, this paper has a modest goal of providing ways to work towards this ideal scenario.

Currently search engines exist that attempt to provide this functionality, such as MEDLINE, but they require extensive human interaction for tagging the relevant information as compared to a more automated one that is discussed in this paper, and therefore are less practical and unable to keep up with the ever increasing medical data.

For the purpose of this paper a single medical information library, MEDLINE, is used. It can be extended to incorporate other sources. MEDLINE uses MeSH (controlled vocabulary) to mark the documents. We would also use UMLS (Unified Medical Language System) for identifying the semantic types of the medical concepts. Table 1 defines the search attributes.

Table 1: Formulae for search attributes [1].

		Manual Review	
		Meets Criteria	Does not meet criteria
Search Terms	Detected	a	b
	Not Detected	c	d
		a+c	b+d

Sensitivity = a/(a+c)
Specificity = d/(b+d)
Precision = a/(a+b)
Accuracy = (a+d)/(a+b+c+d)

2. EXISTING COMPONENTS

Mentioned below are some of the components that are going to be utilized in this work. It is important to remember that this work is really an amalgamation of various existing techniques that have been proven to be very useful over a period of time and that such a task has not been undertaken as yet and therefore no comparisons can be drawn from previous such related works other then the fact that there has been research in the fields of semantic web which deals with the global aspect and requires a revamp of current resources rather then actually using the existing technologies to leverage the newer methodology.

2.1 UMLS

Initiated in 1986 by Donald Lindberg, the Director of the Library of Medicine, it is a collection of controlled medical vocabularies. Besides acting as a repository, it also provides a mapping structure between different terminologies. This paper looks at two parts of the UMLS:

i. Metathesaurus (a collection of concepts and terms from the various controlled vocabularies and their relationships).

ii. Semantic Network (set of relationships among the elements or concepts in the Metathesaurus).

For the purpose of this paper, we will also assume that the Metathesaurus only contains the MeSH vocabulary and it will be organized according to concepts name with a finite set of attributes to define the concepts' meaning.

Semantic networks assign semantic types to each concept in the thesaurus and will show the list of possible relationships between semantic types (e.g. penicillin is an antibiotic, antibiotic being a high level concept according to the MeSH terminology)

2.2 MeSH

Created by the National Library of Medicine, MeSH [7, 8] is a hierarchically, controlled vocabulary for indexing medical documents and journals

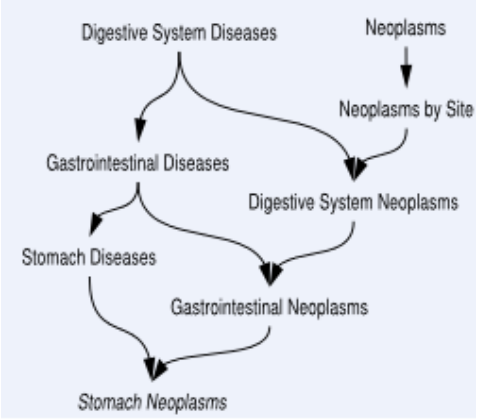

Figure 1: Sample MeSH snapshot [8]

3. OPTIMIZING INPUT

To optimize the input, two techniques are used, namely generic queries and classification.

3.1 Generic Queries formulation [2]

Following an assumption that the user queries can be mapped to a finite set of 'generic queries', we can optimize the input by creating a generic set of predefined queries. A set of user queries are used and semantically analyzed to determine 3 things:

i. Meaning (medical concept),
ii. Generic query template and
iii. Information source.

Generic queries are developed by using a set of databases that contain user queries. Experienced librarians analyze these queries and come up with generic queries using knowledge acquisition techniques to determine the nature of question and the relation between concepts. Results are then placed in the Metathesaurus. These results contain the concepts, which are changed into their corresponding semantic types, as well as their underlying relationship. The architecture for generic query formulation id shown in figure-2.

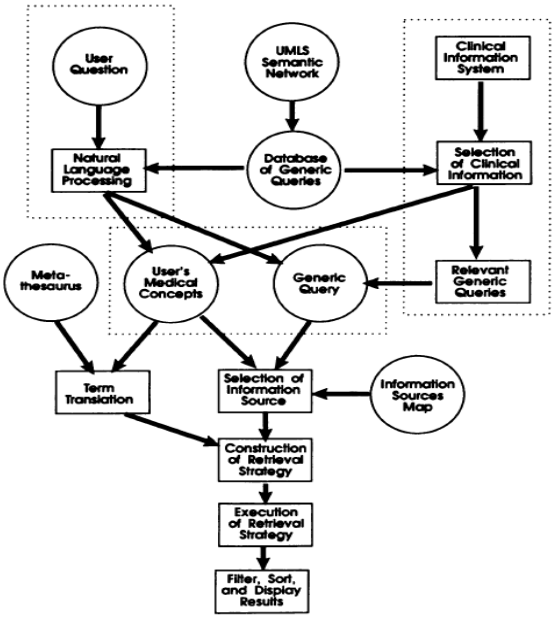

Figure 2: Architecture for Generic query formulation [2]

The information sources also have particular ways that optimum information can be accessed. These "commands" are also mapped to the generic queries indicating that once a generic query is found, it is easily transformed into that command.

Now when a user submits a query, e.g. "Is aspirin the best treatment for Bouillaud Disease?", the semantic analysis yields that the medical concepts ("aspirin" and "Bouillaud Disease") can be mapped to the generic query for "therapy effectiveness". Metathesaurus is

used to come up with the semantic types (pharmacological substance for aspirin and disease for headache). The query template for this user query might look something like <pharmacological substance><treats><disease>. Note that many queries can map to this criteria and so querying using this would yield numerous false results. After finding this generic template, a retrieval command is generated based on MEDLINE syntax. For this example, it can be something like <X> treats the disease <Y>. Earlier we had already identified the medical concepts in the query. Looking them up in MeSH, we find that aspirin remains the same but Bouillaud Disease is tagged as Rheumatic heart disease. So the final MEDLINE query becomes : <Aspirin> treats the disease <Rheumatic heart disease>

3.2 Quality of generic queries

Many studies have tried to shown that queries that use specific words from the vocabulary of the information source tend to produce very accurate results (such as [3]). Work has been done by Wanda Pratt [5] and Haynes [1] as they have classified queries according to major categories.

Wanda [5] discusses how to classify the user queries, without user intervention, according to the content they present. There are 2 steps to categorize queries: lexical analysis and semantic analysis. Because we have already simplified our queries to the smallest possible, we can safely eliminate one of the phases. We can simply look for the particular semantic relationship word and then look up different phrases that we can append and even append all available, e.g. <Aspirin> treats the disease <Rheumatic heart disease>, we know that we can use the corresponding words ("treatment" in this case) for treatments to find the studies.

Table 2: Purpose categories and their criteria [4] [5] [1]

Purpose	Criterion
Etiology	Formal control group: random or quasi-random allocation of participants to treatment and control trial
Prognosis	A cohort of subjects who have the disease in question at baseline without the outcome of interest
Diagnosis	Provision of sufficient data to calculate the sensitivity and specificity
Treatment	Random allocation of participants to treatments
Review	Reproducible description of the methods for conducting the review

4. OUTPUT OPTIMIZATION

The work of Wanda Pratt [4] could be used to optimize the output such that the user can get to the needed information much more quickly. The paper combines the benefits of clustering with the main benefits of classification techniques. The technique is called dynamic categorization as it categorizes the documents/results retrieved under dynamically generated labels.

Using clustering and classification techniques have their own problems. Clustering usually groups documents based more on their structure (so that e.g. documents with similar word count may be placed in the same category) rather then having meaningful labels depending on the user query [6].

Classification allows for having more meaningful labels but the task has to be done manually and is very time consuming. It also presents a problem that every user has a different perception about a single piece of information therefore the labeler would have to account for all the dimensions.

DynaCat has the benefits of both the techniques i.e. it makes meaningful labels that are relevant to the user query. The author [4] claims that such a technique will provide information about:

i. What kind of information is represented in the list entries?

ii. How the documents relate to the query?

iii. How the document relate to each other?

It requires two models: terminology model and query model. The terminology model is the same that we have been assuming throughout, i.e. MeSH keywords and UMLS semantic types.

The query model will provide information about how and what types of queries the user will make, what categories apply to those queries and what is the criteria to look for in those queries. We have already defined the generic queries that are formed during the input. The query type is mapped to the categorization criteria and a label generator that defines how to generate labels. The categorization criteria would be a set of

semantic types for MeSH keywords and subheadings and the label generator will return the appropriate word that fits into those.

During the document look up, each MeSH term in the document is traversed and compared with the categorization criteria. Upon finding a match with the semantic type and the subheading, a label is generated for that MeSH term if not already made.

If the forthcoming results are either too deep or wide, they will still pose significant problem to the user, so another component, the organizer, handles this. For results that are too broad, it traverses up the MeSH tree until it comes up with a parent that is able to cover some of the categories and then generates that label and places other corresponding categories under it.

5. AGENTS

This whole task appears very cumbersome and for the user to actually get the results or suggestions explicitly without their interventions, agents have to be utilized. But before deciding how agents can be of any use, let us first look at a general classification of agents according to their intended behavior.

5.1 Typology [9]

Researchers have various views and sets of characteristics to try to classify agents. They can be classified according to their ability to move around a network thus they can be static or mobile. They can be deliberative or reactive based on whether they are engaged proactively in planning and negotiations or just respond to a certain stimulus. A classification that is most usually employed and widely agreed upon is based upon the characteristics that agents should exhibit: Autonomy (independence), learning (intelligence) and cooperation, as depicted in figure 3.

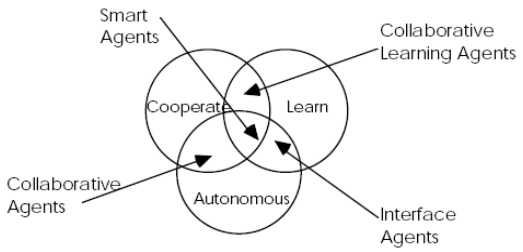

Figure 3: "A part view of an Agent Typology" [9]

5.2 Where Agents fit into our work?

We are going to be using interface agents for our work. But it is imperative to understand that the topology in figure 3 does not necessarily mean that interface agents have nothing to do with cooperation. In fact it depicts that their primary concern is learning and acting autonomously.

The agent would be working as a personal assistant to the physician and just like any assistant needs training, the agent will gradually learn the habits and preferences of the physician.

Pattie Maes [10] describes how agents can simulate their human counterparts in that they possess a minimal amount of information initially and gradually learn from the user depending upon how users perform their tasks. There are 4 different approaches to learning [10].

i. "Learning over the shoulder", i.e. looking for patterns in the user behavior and then trying to replicate them for the user.
ii. User feedback.
iii. Examples given by the users to train the agent.
iv. Collaborating with other agents to know how they perform in certain tasks.

While some prototypical agents have been created to prove these concepts, these are still in early stages. However, assuming that we have access to such an agent, the physician will be able to train the agent better.

5.3 Clinical Agents

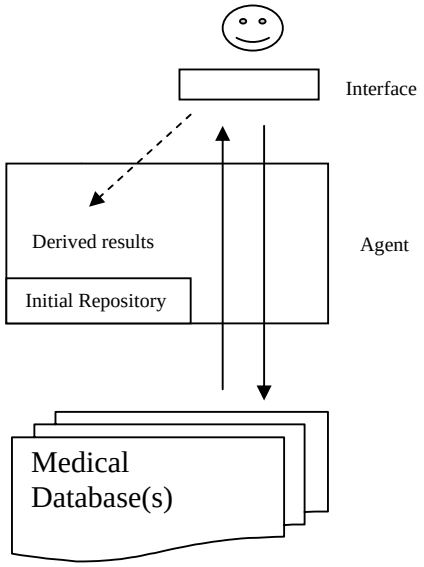

Figure 4: Proposed architecture of an agent

The agents to be utilized would fall into the category of collaborative learning agents. Each physician would have a personal agent to his/her disposal. Initially the agent will possess some information about the medical domain, but as the user queries MEDLINE, the agent memorizes patterns in the user queries (figure 4). This would require that the agents to be equipped with the methodology described in this paper. After getting the results back from the medical database, the user then goes on to select a particular one from the many categories available. The agent thus makes a connection between what the user gave as an input and what he wanted as an output. These patterns would keep building up and after a passage of time and some false positives, the agent can successfully predict what the user wants and come up with more accurate results and suggestions. Because these would be collaborative agents, they can also seek assistance from agents that work for other physicians and come up with suggestions regarding novel issues. It can also provide statistics, e.g. 10% of orthopedic surgeons prescribed aspirin for juvenile arthritis and 8% percent reported instant relief. This would provide the physician with numbers to rely upon and he/she can make a more sound decision. One very obvious and immediate concern would be the security aspect of the information. The agents would have to make sure that their collaboration is with trusted agents and that the information provided by such agents is indeed correct.

6. CONCLUSION

Currently, an enormous amount of data is stored in the medical libraries all over the world but its real power cannot be utilized due to the fact that it cannot be retrieved in a timely manner. I have made a modest attempt to combine all the well-known techniques for providing the physician with the relevant data that is required plus a way to use agents to automate the whole process.

REFERENCES

[1] Brian Haynes et al., "Developing Optimal Search Strategies for Detecting Clinically Sound Studies in MEDLINE", *Journal of the American Medical Informatics Association Volume 1 Number 6,* Nov/Dec 1994.

[2] James Cimino et al., "Generic queries for meeting clinical information needs", *Center for Medical Informatics, Department of Medicine, Columbia University College of Physicians and Surgeons, New York 10032.*

[3] Eneida A. Mendonca et al., "Automated Knowledge Extraction from MEDLINE Citations", *Department of Medical Informatics, Columbia University, New York, NY, USA*

[4] Wanda Pratt, "Dynamic Organization of Search Results Using the UMLS", *Proceedings of the American Medical Informatics Association (AMIA) Fall Symposium (Formerly SCAMC).* 1997.

[5] Wanda Pratt, "QueryCat: Automatic Categorization of MEDLINE Queries", *Information and Computer Science Department, University of California, Irvine, USA.*

[6] Wanda Pratt, "The Usefulness of Dynamically Categorizing Search Results", *The Journal of American Medical Informatics Association* Accepted May 3, 2000.

[7] "MESH", http://www.nlm.nih.gov/cgi/mesh, visited on 25/09/2007.

[8] "Medical Subject Headings", http://en.wikipedia.org/wiki/Medical_Subject_Headings, visited on 25/09/2007.

[9] Hyacinth S. Nwana, "Software Agents: An Overview", *Knowledge Engineering Review, Vol.*

11, No 3, pp. 205- 244, October/November 1996.

[10] Pattie Maes, "Agents that Reduce Work and Information Overload", *Communications of the ACM. July 1987, Vol.37, No. 7, pp.30-40*