Organizing Literature Information for Clinical Decision Support

Dina Demner-Fushman, Susan E. Hauser, Glenn Ford, George R. Thoma

Communications Engineering Branch, Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland, USA

Abstract

Answers to clinical questions occurring during healthcare practitioner/patient interaction can be often found in National Library of Medicine's (NLM) databases. The recent advances in wireless handheld computers promise to make them a widely used tool to deliver needed information to the practitioner at the point of service. This paper addresses challenges in organizing and presenting information obtained from NLM's MEDLINE® database of indexed citations in a way that will help practitioners reduce literature search time on handheld computers. We study two clustering algorithms and two methods of labeling document clusters.

Keywords:
Handheld Computer; Computer-Assisted Decision Making.

Introduction

Answers to the majority of clinical questions that occur during health practitioner-patient interaction can be found in MEDLINE and other databases. However, point-of-care searching was shown to be not fast enough to answer most of the questions [1, 2]. The average times obtained in the clinical questions answering studies are likely to be even longer for handheld computers (PDAs), since, due to limitations of the small screen, at most five truncated citation titles can be displayed on one PDA screen. This means that users who want to read the titles of more than five citations will need to perform several stylus actions. In experiments aimed at reducing access times for information search tasks and number of stylus actions on handheld computers, single screen summaries of web pages have been found effective [3]. Compression of retrieval results into a single PDA screen requires prior organization of retrieved citations. Clustering has been recently reconsidered and shown effective as one of the methods of organizing and presenting relevant documents [4, 5]. Our first approach to clustering followed the bibliographic tradition of classifying items into subject areas [6]. An alternative approach is to cluster documents dynamically based on similarities between documents retrieved for each query. In the latter case a question of naming clusters arises. Selecting the name from the set of MeSH terms assigned to retrieved citations, generating names using controlled vocabularies, and generating names for each cluster from the search results within the cluster are some of the proposed solutions for the cluster-labeling problem [7, 8, 9].

Our goal is to explore methods that can provide a compact single PDA screen overview of retrieved citations, and the possibility of immediate access to relevant documents in the context of the PubMed on Tap application [10].

This paper describes our experiments with both clustering approaches: Subject Area classification of documents using a constant set of pre-determined categories, and Dynamic Clustering using hierarchical clustering methods. Dynamic Clustering requires generation of multi-document cluster labels. We considered two methods of cluster name generation: extraction of multi-document summaries, and selection of the most representative title from the set of citation titles in the cluster.

Methods

At the first stage of our experiment citations for a query are retrieved using PubMed, NLM's interface to the MEDLINE database. These citations are then used in clustering and labeling experiments.

Queries

Since our goal is to organize documents retrieved by PubMed in response to a clinical query, we created our query set by combining several of the 106 OHSUMED collection queries generated by clinically active physicians and reference librarians who regularly used MEDLINE [11], and queries that occurred in point-of-service situations involving the first author. In addition to queries, this set contains descriptions of clinical situations that caused the health practitioner to seek additional information. Queries were selected based on the total number of citations retrieved by PubMed for each query. The number of documents retrieved for each query in the full set of queries ranged from 0 to 14,684. Queries with document counts between 0 and 10 were removed from further consideration, since we do not plan to cluster documents retrieved for focused searches that result in small numbers of relevant citations. The remaining 50 queries were used to

study the distribution of documents clustered using different approaches. A subset of these queries, where document counts did not exceed 70, was studied more closely to evaluate the quality of created clusters and cluster names. The upper bound on the size of the retrieved documents is established to permit exhaustive evaluation of the summaries.

Clustering

1. Subject Area clustering

We use the *Alphabetic listing by subject field* section of the list of 4,500+ journals being indexed for Index Medicus® [12] as the main source of subject areas. Unfortunately, this section contains only about a third of the 10,192 *Serials Indexed for Online Users*. We augmented the list of subject areas with the set of controlled descriptors used for indexing journals according to discipline in NLM's Indexing Initiative [13], thus obtaining subject areas for 8,764 journals. Previous approaches that used controlled vocabulary to generate subject area names used one of the unique vocabulary terms to name each of the areas. In these approaches each document is placed into as many subject areas as it belongs, judging by the journal that published the document. We chose to create new subject areas that combine all single controlled vocabulary terms that are assigned to the journal for two reasons: 1) in many cases one of the subject areas is very broad, and the remaining subject names clarify which aspect of the broader area is the closest to journal's publications; 2) a significant reduction in space required to display citations' titles belonging to each subject area, which is particularly important because of the PDA's small screen. As an additional benefit, this approach permits creation of a hierarchical structure using the merged cluster names, as for example in the case of Allergy and Immunology [Figure 1]. This may help reduce even further the initial space requirements to display top-level subject areas. For example, 42.5% of the citations for the query "*Catamenorrheal Anaphylaxis*" are in the Allergy and Immunology area. We can reduce the size of this area by reassigning approximately half of citations to three merged areas: *Allergy and Immunology; Pulmonary Disease (Specialty), Allergy and Immunology; Parasitology,* and *Allergy and Immunology; Pediatrics.* The size of the top-level display does not increase, since it still contains only the Allergy and Immunology area name. The Subject Area clustering method classifies the 8,764 journals into one of the 1,332 unique subject areas. The number of journals in each area varies, e.g., *Tuberculosis* subject area contains six journals, but 659 journals belong to the subject area *Medicine.*

2. Dynamic Clustering

Dynamic Clustering requires determining the similarity between documents. We compute the similarity between documents using the traditional vector space model [6], where each document is represented as a vector of terms that occur within the document collection (MEDLINE in our experiments). The importance, or weight of each term found in document *d* is computed as

$$w_i = tf_i \cdot idf_i \tag{1}$$

where tf_i is the frequency of the ith term in the document d, and idf_i is the log of the term's inverse document frequency in the document collection computed as

$$idf_i = \log_2\left(\frac{n}{df_i}\right) \tag{2}$$

where n is the total number of the documents in the collection, and df_i is number of the documents that contain the ith term.

Similarity between two documents is computed as the cosine of the angle between corresponding document vectors in the N-dimensional space, where N is the number of terms in the document collection

$$sim(\vec{d}_j, \vec{d}_k) = \frac{\sum_{i=1}^{N} w_{i,j} \times w_{i,k}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{N} w_{i,k}^2}} \tag{3}$$

We then use the Ward's hierarchical clustering algorithm [14] to create document clusters. The number of clusters in hierarchical clustering depends on the similarity threshold. This threshold defines when two documents are no longer considered similar enough to be placed in the same cluster.

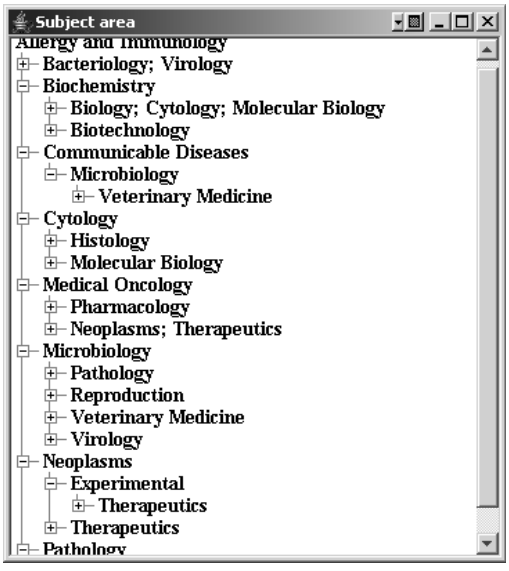

Figure 1 - Subject areas within Allergy and Immunology.

Dynamic Cluster labeling

The recent Genomics track TREC experiments at NLM demonstrate that in many cases the title of the MEDLINE citation is the most representative sentence for this citation [15]. The experiments aimed at reducing search time for information in web pages, described in [3], show the most significant time reduction when the representative single sentences combined with the most salient terms in the web page text are used as the summary of the web page. Based on this knowledge we decided to test two alternative approaches to cluster name generation. We used the multi-document extraction summarization tool developed at USC/ISI [16] to obtain extractive summaries as the labels for each cluster. As an alternative we selected the most representative title from the set of titles in each cluster as follows: the normalized weight was computed for each title as shown in Equation (4):

$$w(t) = \frac{\sum_{i=1}^{n} w_{i,t}}{n} \qquad (4)$$

where n is the number of documents in the cluster, and $w_{i,t}$ is the weight of the terms in the title. The title with the highest weight was selected to represent the cluster.

Evaluation

We chose to use intrinsic (normative) evaluation of information organization at this stage of development of our application [17]. This evaluation is usually based on user judgments. We evaluated partitioning of the documents into clusters and cluster names as follows: given the cluster names, the query and the clinical situation described in metadata for the query, select the cluster that most probably contains the answer to the query. Retrieval results partitioning was considered successful if the first selected cluster provided an answer to the query and the cluster label reflected the main topic of each citation. This evaluation is suitable for the early formative design stages. We plan a rigorous assessment at the end of the development cycle with the emphasis on task-oriented methods that measure the users ability to perform specific tasks [18].

Discussion

Subject Area clustering

Classification by Subject Areas is very appealing because it does not require additional processing: the pre-defined categories are looked up in a table while each citation is prepared to be sent to the client. In some cases this is sufficient to direct a user's search in the right direction. For example, citations retrieved for the query *"trigeminal neuralgia combination drug treatment"* were partitioned by this method into nine Subject Areas. The area named "Drug Therapy; Pharmacology; Pharmacy; Therapeutics" was selected as the most promising, and the answer was found in one of the five citations assigned to this cluster [Figure 2].

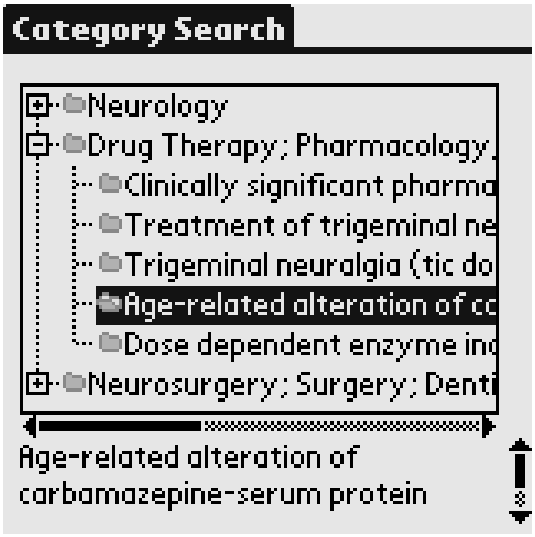

Figure 2 - PubMed on Tap search results for the query "trigeminal neuralgia combination drug treatment" organized by category. The collapsed tree permits users to select the most promising category without scrolling. One pen action provides the overview of the titles. The highlighted title is displayed in the lower part of the screen in more detail

A drawback of this approach is in uneven and unpredictable distribution of documents between categories for individual queries. Another drawback is a fairly large number of not categorized journals. Results from eighteen of the fifty searches contained documents that could not be placed in one of the subject areas. On average, 11.74% citations in the eighteen result sets were labeled as not categorized [Figure 3].

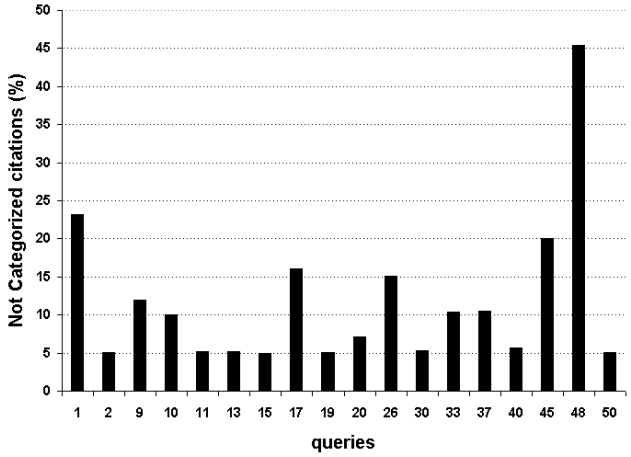

Figure 3 - Percentage of Not Categorized citations in 50 sets clustered using Subject Areas

The uneven distribution of citations results from the journal profiles, since many journals, for example *JAMA*, cover broad areas. In addition, areas represented by only one category in the subject listing, for example *Veterinary Medicine,* contribute to uneven distribution. Therefore it is not surprising that we obtained a wide spectrum of distribution patterns among documents retrieved for 50 queries selected to evaluate the

Subject Area clustering method. Figure 4 presents a typical distribution pattern. More than half of the retrieved documents are in categories *Gastroenterology* and *Surgery*. The rest of the documents are distributed relatively evenly between the remaining eight categories. There was no correlation between the number of retrieved documents and the number of subject categories for each query (ρ=0.029)

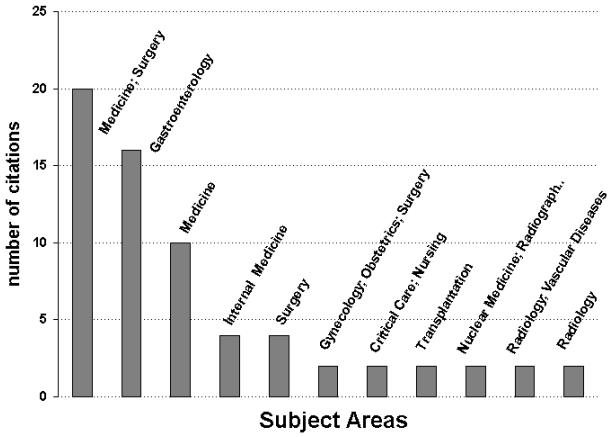

Figure 4 - Distribution of citations for one of the queries in the OHSUMED collection (query 73: "portal hypertension and varices, management with TIPS procedure")

Dynamic Clustering

In the preliminary evaluation of the dynamic organization of retrieval results we modeled a clinical scenario based on the first author's experience. The clinical situation involves a female patient with the history of good oral health, who started hormone replacement therapy and presented symptoms of gingivitis. Patient's dentist is interested if there is a relationship between hormone replacement therapy and gingivitis. The dentist is also looking for recommendations for gingivitis management in this situation. The Subject Area clustering was not very helpful in organizing the results for the query formulated as follows: "*gingivitis hormone therapy"*. When Subject Area clustering was used, 53.8% of the citations fell into the *Dentistry* category, and another 10.3% could not be categorized. Dynamic Clustering, on the other hand helped select the relevant cluster judged by the best title name "*Female hormones and oral health*". We considered several similarity thresholds while clustering citations retrieved for this query, and selected one that resulted in four clusters. This partitioning was chosen because the citations were divided into distinct areas and the number of citations in each cluster was manageable. The extractive multi-document summaries for the clusters were not always meaningful. They also did not always reflect the main topic of the citations in the cluster, for example the first cluster in Table 1 was summarized by the first author as follows: *Systemic diseases and conditions that predispose an individual to periodontal destruction are described. One of the papers discusses differential diagnosis of gingivitis in children*.

Table 1 - Cluster name generated using two labeling methods.

Multi-document extractive summary, extracted phrases are slash-separated	Best title	Number of citations
CHILDREN INCREASING INFORMATION EMPHASIZE / EARLY TREATMENT OF PERIODONTAL DISEASES / ORDER TO AVOID ERRONEOUS / PEDIATRIC DENTIST / AGE PRESENT REVIEW OUTLINES STRUCTURAL	Periodontal manifestations of systemic disease	4
SYSTEMIC DISORDERS HAS PROMPTED RESEARCHERS TO INVESTIGATE / AUTHORS / HORMONAL CHANGES / PREGNANCY OUTCOMES CARDIOVASCULAR DISEASE / OSTEOPOROSIS METHODS / HORMONE LEVELS	Desquamative gingivitis: revisited	17
SEVERITY AND PROGRESSION OF PERIODONTAL DISEASE WE SOUGHT TO FURTHER EXPLORE BIOLOGIC / RENDERED DIABETIC BY 65 MG INTRAVENOUS INJECTION	Diabetes mellitus and periodontal disease	6
FEMALE HORMONES AND ORAL HEALTH COMMON ORAL MANIFESTATION OF ELEVATED LEVELS OF OVARIAN HORMONES ESTROGEN AND PROGESTERONE AS SEEN IN	Female hormones and oral health	7

Conclusions and future work

Our preliminary results show that it is possible to direct a user's search towards the most relevant documents. We were able to quickly answer clinical queries using both methods of partitioning of the retrieved documents. Categorization by Subject Areas was not always helpful in directing user's search. In the worst cases the most promising subject areas were too general and contained the majority of the documents. Dynamic Clustering and cluster labeling using the best title show promise, but need further exploration of the computational requirements for large sets of retrieved documents, and dynamic threshold determination. We plan to measure time requirements for both methods, evaluate clusters obtained using different threshold values in the Dynamic Clustering approach, determine percentage of unsatisfactory distribution patterns in the Subject Areas partitioning approach, and use task-oriented controlled experiments and traditional information retrieval performance measures to evaluate clustering results.

Acknowledgments

Authors would like to thank Susanne Humphrey for valuable discussions of NLM controlled vocabularies, and Eduard Hovy, Liang Zhou and Anton Leuski for the opportunity to use their applications.

References

[1] Alper BS, Stevermer JJ, White DS, Ewigman BG. Answering family physicians' clinical questions using electronic medical databases. *J Fam Pract*. 2001 Nov; 50 (11): 960-5.

[2] Chambliss ML, Conley J**.** Answering clinical questions. *J Fam Pract*. 1996 Aug;43 (2): 140-4.

[3] Buyukkokten O, Kaljuvee O, Garcia-Molina H, Paepcke A, Winograd T. Efficient web browsing on handheld devices using page and form summarization. TOIS 2002 20 (1): 82-115.

[4] Hearst MA, Pedersen JO. Reexamining the Cluster Hypothesis: Scatter/Gatherer on Retrieval Results. Proceedings of ACM SIGIR '96, Aug, 1996, 76-84.

[5] Leuski A. Evaluating document clustering for interactive information retrieval. Proceedings of Tenth International Conference on Information and Knowledge Management (CIKM'01). Nov, 2001, 41-48.

[6] Salton, G., "The Smart Retrieval System", Prentlce-Hall, Englewood Cliffs, New Jersey, 1971.

[7] Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, Rindflesch TC, Wilbur WJ. The NLM Indexing Initiative. Proceedings of AMIA Annual Symposium *2000***.**

[8] Pratt W. Dynamic Organization of Search Results Using the *UMLS* Proceedings of the American Medical Informatics Association (AMIA). 1997

[9] Vivisimo clustering engine Available from: http://vivisimo.com/demos/PubMed@NIH.html

[10] Hauser SE, Ford G, Demner-Fushman D, Thoma GR. Organizing literature information for clinical decision support. Submitted to: 11th World Congress on Medical Informatics; 2004 Sep 7-11; San Francisco CA, USA.

[11] Hersh WR, Buckley C, Leone TJ, Hickam DH, OHSUMED: An interactive retrieval evaluation and new large test collection for research, Proceedings of the 17th Annual ACM SIGIR Conference, 1994, 192-201

[12] List of Journals Indexed in Index Medicus Available from: http://www.nlm.nih.gov/tsd/serials/lji.html

[13] Humphrey SM., Rindflesch TC, Aronson AR. Automatic Indexing by Discipline and High-Level Categories:Methodology and Potential Applications 11th ASIST SIG/CR Classif Res Workshop, Chicago IL, 12 Nov 2000

[14] Ward, J. H. Hierarchical Grouping to Optimize an Objective Function *J. Am. Statist. Assoc.* 1963, 58, 236-244

[15] Kayaalp M, Aronson AR, Humphrey SM, Ide NC, Tanabe LK, Smith LH, Demner D, Loane RR, Mork JG, and Bodenrider O. Methods for accurate retrieval of MEDLINE citations in functional genomics. Text Retrieval Conference, 2003.

[16] Zhou L, Hovy E. A Web-Trained Extraction Summarization System. Proceedings of the HLT-NAACL conference, Edmonton, May 2003

[17] Mani I, Firmin T, House D, Chrzanowski M, Klein G, Hirschman L, Sundheim B, and Obrst L. 1998. The TIPSTER SUMMAC Text Summarization Evaluation: Final Report. MITRE Technical Report MTR 98W0000138. McLean, VA: The MITRE Corporation.

[18] Hersh WR, Detmer WM and Frisse M.E. Information Retrieval Systems. In Edward H. Shortliffe and Leslie E. Perreault (Eds.), Medical Informatics: Computer Applications in Health Care and Biomedicine (pp. 539-572). 2nd ed. New York: Springer, 2001.

Address for correspondence

Dina Demner-Fushman, Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, Maryland 20894, USA
E-mail: ddemner@mail.nih.gov