
Extending PubMed on Tap by means of MultiDocument Summarization

José Carlos Cortizo¹, Diego Gachet¹, Manuel de Buenaga¹, Manuel Maña²,
Enrique Puertas¹, Manuel de la Villa²

¹ Grupo de Sistemas Inteligentes
Universidad Europea de Madrid
C/Tajo s/n, 28670, Villaviciosa de Odón, Madrid, Spain
josecarlos.cortizo@uem.es;gachet@uem.es;buenaga@uem.es;enrique.puertas@uem.es
² Departamento de Ingeniería Electrónica, Sistemas Informáticos y Automática
Universidad de Huelva
manuel.mana@dti.uhu.es;manuel.villa@dti.uhu.es

Abstract. Access to biomedical databases from pockets and hand-held or tablet computers is a useful tool for health care professionals. PubMed on Tap is the standard application for PDA to retrieve information from Medline, the most important and consulted bibliographical database in the biomedical domain. In this paper we present a description of an intelligent information retrieval system that uses clustering and multidocument summarization techniques improving aspects of PubMed on Tap.

Keywords: ubiquitous computing, medical domain, multi-document summarization.

1 Introduction

Information overload is common nowadays in our society. This is also the case for medical information, available from a variety of sources, including scientific papers, databases of summaries, structured or semi-structured data-bases, web services and clinical records of patients. In this domain professional in general need tools oriented to provide facilities for accessing and visualizing the adequate information for their needs.

In general, work at hospitals requires mobility and coordination due to the complexity of the tasks, intensity of information exchange, and the way information and resources are distributed. Hospital's staff might be distributed in space (i.e. different location within the settings) or time (i.e. working different shifts).

Hand-held computers or PDAs wirelessly connected to a Hospital Information system (HIS) can give physicians access to patient medical records or any relevant related information anytime, from anywhere within the hospital, allowing physicians to take more informed and evidence based decisions in the same patient's point of care.

Mobile devices, even with their limited screen size, offer clear advantages in different scenarios, but the capability to select the key information, and display it

in a synthetic way plays a key role. Given the number and diversity of medical information sources, different methods must have been found that enable users to quickly assimilate the content of a document in the reduced display of a typical mobile device.

In this paper we present a multi-document summarization approach integrated with text clustering useful in this kind of setting. We present the application of our system improving aspects of PubMed on Tap [6]. PubMed on Tap is the standard application for PDA to retrieve information from Medline, the most important and consulted bibliographical database in the biomedical domain.

2 Uses of PDA in the medical Domain

Healthcare providers, especially younger clinicians, residents and medical students, are increasingly adopting the use of Personal Digital Assistant (PDA) for access to a variety of information sources as for example Medline, this is particularly true if we consider that the practice of evidence based medicine (EBM) has grown dramatically since the term was introduced, turning PDAs as a valuable tool for providing information nuggets in a just-in-time manner, the most important tool in this case is PubMed on Tap, a PDA application for accessing on-line Medline information[6]. In the next subsections we analyze the importance of PDA from different points of view.

2.1 Importance of PDA in Medicine

Medical personnel at hospitals need to move continuously around the premises to access people, knowledge, and resources in order to perform their work effectively [7]. Thus, mobility characterizes work in these environments. For instance, physicians make daily rounds to assess and diagnose patients, changing their location to find colleagues or locate artifacts (patient records, X-ray images, medications) placed in bed wards, laboratories or offices.

These challenges are motivating the widespread adoption of Personal Digital Assistants (PDA) computers in support of hospital work. To date, the most popular handheld medical applications are the ones that provide access to reference material, such as pharmacological databases [8] or bibliographic material as Harrison's book of Medicine or medical calculators as Medcalc.

In addition, there is a tendency to use PDAs wirelessly connected to a hospital information system, where this can give physicians access to patient medical records or last evidence information [9]. Even with their limited screen size there are clear advantages from having this increased availability of information. A more recent trend in supporting mobile hospital workers includes the development of pervasive computing environments and location-aware information systems [10]. These efforts aim at providing hospital staff with access to relevant information from anywhere within the hospital and on a variety of heterogeneous devices.

2.2 Relevance of Scientific Documentation

The output of biomedical research in the form of literature written in free-form text format (unstructured format unsuitable for complex searching) are then accumulated in large online databases which have been readily accessible due to recent advances in software and communications. In the last years, the Web has changed the way information is disseminated, knowledge is gained, and health care is provided.

However, the fast growing of research results from biomedical domain is producing a major bottleneck, an information overload. MEDLINE (Medical Literature Analysis and Retrieval System Online), the U.S. National Library of Medicine's premier bibliographic database, contains over 16 million references to journal articles in life sciences with a concentration on biomedicine. Between 2,000-4,000 completed references are added each day, over 670,000 total added in 2007 [4].

The practice of evidence-based medicine has traditionally been defined as combining the best medical research findings with clinical judgment, expertise and experience. The ability to search the medical literature in a time efficient manner represents an important part of an evidence based practice. The use of electronic databases of pre-appraised evidence can greatly expedite the search for high quality evidence. A recent qualitative study found that two of the six obstacles to answering clinical questions with evidence were the time required to find information and the difficulty in selecting an optimal search strategy [5]. That's why search tools as PubMed¹, BioMed Central² or UpToDate³ have become more and more important, to find efficient ways to locate the best evidence in a timely manner.

2.3 Importance of Scientific Information in PDAs

As we have said before, PDAs find many applications in health care. Family physicians and specialists have been using PDAs for general medical reference, such as drug interactions, pharmacopeias, and cardiac risk. Other important applications of PDAs are those involving data collection and management, as in patient tracking, electronic Case Report Forms in clinical trials, patient diaries, and infection surveillance.

To consider the suitability of PDAs accessing information and across all health care contexts, it's interesting to consider some patterns of using data. A recent review about estimates of PDA use by health care providers [11], shows an evolution of PDA usage by health care providers ranging from 30% in 2000 to 60% in 2006. The review analyzes several patterns of PDA usage considering age, students and medical residents, gender, family physicians versus specialists, and large and hospital-based practices. In the professional use, the information access

¹ <http://www.nlm.nih.gov/pubs/factsheets/pubmed.html>

² <http://www.biomedcentral.com/info/>

³ <http://www.uptodate.com/home/about/index.html>

related to drug information is the one presenting more frequent pattern (93%) in terms of patient care, above other such as patient records (43%) or medical calculator (43%).

PDA based information access solutions are developed, in general, for mobile health care professionals who seek medical information when away from their desktop computers. Those solutions are oriented to several scenarios including the point of care. Finding the best evidence to answer clinical questions is one of the basic steps in evidence-based medicine (EBM) practice, and to be most effective, the practice of EBM must occur in real-time at the point of care: physicians almost never seek answers to clinical questions after the clinical sessions ends [12]. Different resources may be valuable in that scenario. In relation with the more specific elements in which we have focused our work, a recent study [6] concludes that handheld computers with Internet access are useful tools for healthcare providers to access Medline in real time, and Medline citations can answer specific clinical questions when several medical terms are used to form a query in the point of care using PubMed on Tap.

3 Extending PubMed on Tap

Medics usually cannot consult big documents when they are visiting patients. Not only because of the low time, but also because the patient could think they don't know what they have if they see the medic consulting information in front of them. This does not mean that medics should not consult information. They have always done this, using traditionally a pocket notepad, but the kind of information and the way it is consulted is very different from what they do when they are not with the patient.

So if we want medics to use PDAs as a substitute of the traditional notepad, we have to show information in a different way, smaller text documents and having an advanced interface made having in mind usability and efficiency so they can make queries and find the information very quick.

To do so, we propose to make clusters of documents and summarize them in order to show to the user only the summaries of each cluster. With this approach we are reducing the number of shown results and the length of the documents, and this fits very well in a system running in a PDA or any other mobile device with advanced capabilities.

This opens a wide range of use cases. Before, the medic would not consult information when visiting patients because time to find what he wanted to find could be very long, but now, with a simple query he can see in the summary all the relevant information. This is useful in atypical cases when the treatment is not very obvious for the doctor.

4 System Architecture and Design

From the system architecture point of view, the PDA communicates with a server. We choose this approach due to limited processing power of PDA and then the whole processing is performed at server side. The client part running on PDA only needs to run a web browser that will execute the web application that will allow the user to send queries and present results.

As we can see in the subfigure 1.A, within the server there are running some processes, the most important are a servlet that receive queries from the PDA and translates back the results obtained, and the multi-document summarization system which mission is clustering and segment documents.

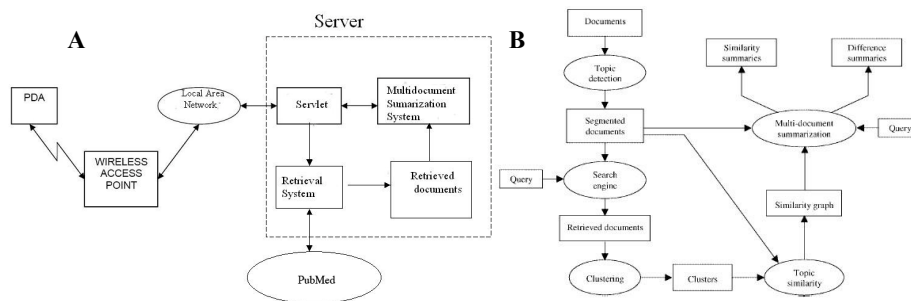


Fig. 1. Subfigure A show the system diagram with main components. Subfigure B shows the architecture of the multidocument summarization system

A more detailed diagram of the multi-document summarization system is presented in subfigure A.2. In this case documents are segmented into multi-paragraph units, reflecting their subtopic structure. Query results are clustered by semantic closeness and a graph that maps the similarity relationships among topics for each cluster is obtained. Information about document topics in each cluster and their similarity is used to compute a summary for the cluster central topic and also summaries focused on particularities of each document.

The software client running on PDA only transmits to the server the necessary data for the query and waits for the results in order to be displayed on the PDA screen.

For preserving platform independence we decided to use a web application specifically designed to show the results on a PDA or a mobile device. The only requirement for running the web applications is the availability of a browser and connection to the network in the mobile device. In order to take the most of the screen, we detect the resolution of the mobile device and the server formats the results according to the resolution of the screen, showing more results on larger displays.

This approach presents some challenges, as designing the web application to minimize the information exchanged between server and client, choosing how to represent the results according to the resolution of the screen and designing a usable interface for the web application.

5 Multi-document Summarization

Text summarization could play a relevant role in the access to the required information at the right time and in the most suitable form both in scientific and clinical practice. Potential applications include teaching and learning, continuing professional education, and improved patient care [1]. In [2] five reasons are argued to generate automatic summaries even having an author's abstract.

A system that exploits the capacity of organization of clustering techniques but being, also, able to complement groups and documents with useful information, can be of great utility in a mobile retrieval setting. We propose to integrate multi-document summarization (MDS) techniques with a post-retrieval clustering interface. The final result is a system that offers a summary for each cluster reporting document similarities and a summary for each document highlighting the singular aspects that it provides with respect to the common information in the cluster. Results obtained in experiment where users performed interactive searches on PCs have been encouraging [3].

The similarity relationship among topics (i.e., document segments) in a cluster is computed to identify the central topic in this cluster and the original aspects in each one of the documents. This information sets up the basis of summaries focused on similarities and differences, respectively.

5.1 Similarity Summaries

It is expected that a cluster of documents addressing similar contents covers a huge variety of different subtopics in relation to the main subject. Each document of the cluster may address one or several of these aspects making very difficult a summarization without a previous identification of, at least, topic boundaries. After that, it is possible to find semantic relations among segments of different documents and thus, determine which of them define the common story line in the cluster. We propose to exploit only the most significant segment. In this setting, it is not probable a crucial loss of information that hampers the identification of the common idea behind the group of related documents. In contrast, the simplification of MDS problem is evident. The most prominent gain is that reducing to one the number of sources to handle we avoid the problems of redundancy and inconsistency inherent to summaries extracted from different sources.

In order to recognize the common topic in the bag of segments corresponding to a document cluster, the similarities among them are characterized using a graph. Similarities between segments are computed as the inner product of both vectors of term weights (tf · idf type). An edge is set up between the nodes representing

two segments when the similarity between them exceeds certain threshold. Then, for each cluster, we chose the segment with the highest degree in its similarity graph.

Sentence extraction techniques allow constructing summaries that are domain independent. The summaries are generated selecting sentences of the original document that contain information highly indicative of its content. The selection is made scoring each of the sentences using a heuristic set. Finally, the sentences with the highest scores are chosen. Among the most used heuristics for selecting sentences, we have chosen the centroid, title, location and query methods [3].

5.2 Difference Summaries

A summary of each document oriented to highlight the differences with respect to the rest of documents in the cluster is produced. Sentences are scored using heuristics related to the significance of the segment they belong to and to the significance of each sentence within the segment. Previously, a redundancy factor is used to discard sentences that repeat information of other ones contained in the similarity summary.

For our interests, the significance of a segment depends on the originality of information that it contains with respect to the information contributed by the similarity summary. Nevertheless, this significance rate must consider the relevance of the segment regarding the rest of the document. Thus, we attain to avoid that segments with a limited relevance in the document stand out in the summary, trying in this way to balance the novelty of the information and its relative significance in the document.

6 Access Methods and Interface

The graphical user interface is very simple, integrating specific elements taken from intelligent information access techniques, which permitted us to exploit text contents analysis. Although space is at a premium on the small screen, readability trumps space utilization, so icons, white spaces and different colors are used to separate distinct items on the screen.

The access to the system is very simple. The user introduces a query and waits the results to be displayed. An example of results disposition is shown in Figure 3, where each cluster is represented with a plus icon, the name of the cluster, the number of documents within the cluster and a summary based on similarities. The user can select a cluster, to display more clusters or perform a new search. If the user selects one cluster, the system shows the title of the documents and a summary of each document based on the differences (Figure 3, right subfigure). At this point, the user can select a document, to show more documents in the cluster or perform a new search.



Fig. 3. Screenshots showing the results of using the prototype. The left image shows the view of the clusters, right image shows a cluster

7 Future Work

We are developing a full operative version of the system in a controlled environment, and planning to evaluate it over an adequate set of users, including medical researchers and students and a domain of scientific information and clinical cases. Also, we want to adapt elements to conduct a TREC-based adequate evaluation method, as the one in interactive track to analyze in more detail the effectiveness of our approach.

References

1. Afantenos, S.D., Karkaletsis, V., Stamatopoulos, P.: Summarization from Medical Documents: A Survey. *Artificial Intelligence in Medicine*, 33(2):157-177. (2005)
2. Reeve, L.H., Han, H., et al.: Concept Frequency in Biomedical Text Summarization. 15th Conference on Information and Knowledge Management, 604-611. (2006)
3. Maña, M.J., Buenaga, et al.: Multidocument summarization: An added value to clustering in interactive retrieval. *ACM TOIS*, 22:2(215-241). (2004)
4. MEDLINE Factsheet. <http://www.nlm.nih.gov/pubs/factsheets/medline.html>
5. Ely, J.W., Osheroff, J.A., et al.: Obstacles to answering doctors' questions about patient care with evidence: qualitative study. *BMJ* 2002;324: 710 (2002).
6. Hauser, S.E., Demner-Fushman, D., et al.: Using wireless handheld computers to seek information at the point of care: an evaluation by clinicians. *Journal of the American Medical Informatics Association*; 2007 Nov-Dec ; 14(6):807-15
7. Bardram, J.E. and Bossen, C.: Moving to get a head: local mobility and collaborative work, *European Conference on Computer Supported Cooperative Work*, pp.355–374.
8. Lapinsky, S.E., Wax, R., et al.: Prospective evaluation of an internet-linked handheld computer critical care knowledge access system, *Critical Care*, Vol. 8, No. 6.

9. Wang, J. and Du, H.: Setting up a Wireless Local Area Network (WLAN) for a healthcare system, *Int. J. of Electronic Healthcare*, Vol. 1, No. 3, pp.335–348. (2005)
10. Munoz, M.A., Rodriguez, M., et al.: Context-aware mobile communication in hospitals, *IEEE Computer*, Vol. 36, No. 8, pp.38–46.(2002)
11. Garrity, C., El Emam, K.: Who's using PDAs? Estimates of PDA use by health care providers: a systematic review of surveys. *J. of Medical Internet Research*; 8(2) (2006)
12. León, S.A., Fontelo, P., et al.: Evidence-based medicine among internal medicine residents in a community hospital program using smart phones. *BMC Medical Informatics and Decision Making*; 7: 5. (2007)