Procesamiento del Lenguaje Natural, Revista nº 45, septiembre 2010, pp 255-258 recibido 26-06-10 revisado 03-07-10 aceptado 06-07-10

A Biomedical Information Retrieval System based on Clustering for Mobile Devices*

Un Sistema de Recuperación de Información Biomédica en Dispositivos Móviles basado en Agrupamiento

Manuel Millán, Alejandro Muñoz
Escuela Politécnica Superior
Universidad de Huelva
{manuel.millan,
alejandro.munoz}@alu.uhu.es

Manuel de la Villa, Manuel J. Maña
Depto. de Tecnologías de la Información
Universidad de Huelva
{manuel.villa, manuel.mana}@dti.uhu.es

Resumen: La sobrecarga de información producida por la creciente disponibilidad en internet de textos y publicaciones de interés es un problema que se acrecienta cuando esa información es necesaria para la toma de decisiones, como ocurre en el ámbito biomédico. Es en este dominio donde se ubica este sistema de recuperación de información dirigido a dispositivos de consulta móviles, que a los tradicionales procesos de indexado y búsqueda, añade la característica de la devolución de los resultados de manera agrupada en función de su contenido.
Palabras clave: Recuperación de información, dispositivos móviles, agrupamiento de documentos, dominio biomédico.

Abstract: Information overload caused by the increasing availability of online texts and publications of interest is a problem that increases when such information is necessary for decision making, as in the biomedical field. It is in this domain where we present an information retrieval system for mobile devices. Traditional indexing and search processes are enriched with the feature of returning the results in clusters according to their content.
Keywords: Information retrieval, mobile devices, clustering, biomedical domain

1 Introduction

In general, work at hospitals requires mobility and coordination. Hospitals staff might be distributed in space or time and their information needs are highly dependent on contextual conditions (Munoz et al, 2003). For this reason, many hospitals are encouraging the use of PDAs connected via wireless networks to the information systems of hospitals and allowing access to a wide variety of information sources (León et al., 2007). A recent study estimated PDA use by health professionals shows an evolution in the use ranging from 30% in 2000 to 60% in 2006 (Garrity and El Emam, 2006).

In this context it is necessary to make available to physicians and other health professionals, systems that allow access to the information they need from anywhere in the hospital. In this way, doctors could make more informed decisions and based on evidence from the point of patient care, either in a doctor's office or beside their bed.

The post-retrieval clustering is a technique that has been investigating for at least 15 years in order to improve the organization of the results returned by an Information Retrieval System (IRS) and facilitate navigation between them (e.g., Scatter/gather algorithm (Hearst and Pedersen, 1996), (Maña, Buenaga and Gómez, 2004), (Dunlavy et al., 2007).

We propose to integrate the clustering of documents in a SRI in the biomedical domain. Clummed (CLUstering on Mobile MEdical Devices) is the result of this integration. The

* This work has been partially funded by the Spanish Ministry of Science and Innovation and the European Union from the ERDF (TIN2009-14057-C03-03)

system organizes search results into folders that store documents that are semantically related.

The complexity of the information that the user must handle is reduced drastically. Simply reading the names of the folders, the user can choose the group of documents that is closest to their information needs. In a second step, users navigate among the documents of the group, a considerably smaller number of documents that the total recovered, to find those that are of interest.

The paper is organized as follows. In the next section we present the related work presented here. Section 3 describes the proposed system. Finally, in Sections 4 are showed briefly the conclusions.

2 Related work

The biomedical information retrieval systems developed for portable devices appeared relatively recently. Nevertheless, still few systems had taken advantage of accessing information across the Internet or other external real-time information resources.

We can find stand-alone tools, capables of made available data already stored, like medicine's reference guides, clinic practice guides, protocols, algorithms, programs for diverse calculations and e-books.

But the widely extended use is the wireless internet access to perform, besides the previous functions, searches of information on remote resources and communicate with remote clinic information systems, as medical history.

(Hauser et al., 2007) publish a prospective study where results prove MEDLINE, accessed from mobile devices, is a viable information source for supporting clinical decision-making. Currently, the available resources accessible from the PDA at the bedside provided response to 86% of clinical questions, most of them (88.9% - 97.7%) during the rounds of visits.

As already noted, there are several works in other domains that suggest that the grouping or clustering improves the organization of the results returned by an information retrieval system (IRS) but the use of clustering in the biomedical field is a novelty.

In previous work we have presented some prototypes of IRS who made use of clustering of results, as in (Maña, Buenaga and Gómez, 2004) and (Buenaga et al., 2008) where each group is identified by their similarities and each individual document by differences.

3 System overview

Our system follows a traditional pattern, dividing the process in the work of indexing and retrieval. In indexing, a preprocessor prepares and creates the structure that supports the documents, in our case a set of biomedical documents (retrieved from BioMed Central[2], an online repository of biomedical papers mainly open access). With the support of Lucene (Gospodnetic, Hatcher and McCandless, 2009), a library of functions for indexing and searching of texts, written in Java, perform manipulations on the text of each document to obtain a representation of the document in terms or tokens that will form the set of index files.

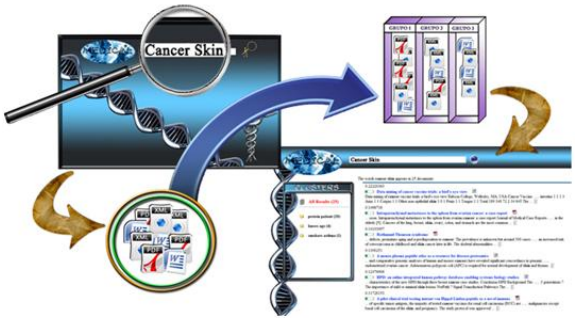

Fig. 1 System operation scheme

The retrieval process (Fig. 1) starts with the input of a search string in the user interface. The search engine will be able to understand and process the query to search in the collection of documents whose index contains some elements of the search string. A similarity algorithm scores the adequacy of each document to the search string and selects a group of documents.

Among the innovations of our system are clustering post-retrieval and the adaptation of the system and its interface for access from mobile devices, showing the user the different groups of documents obtained, among the user navigate, selecting the group he most wants. The system will respond expanding it and listing from high to low relevance the documents included in the group.

3.1 Clustering

Given a set of documents, clustering is the task of dividing the set into groups according to their semantic content. Depending on the structure of groups which produce, clustering

[2] http://www.biomedcentral.com

methods are usually classified as non-hierarchical and hierarchical or partition-based (Rasmussen, 1992).

Methods based on partitions simply divide the set of documents in a number of disjoint groups at the same level, identified by a centroid or center of gravity, representative of the semantic features of the documents it contains. The latter group includes the two clustering algorithms we have worked with, the K-Means (MacQueen, 1966) and the Expectation-Maximization (Dempster, Laird and Rubin, 1977), selected mainly for its availability in the Weka library (Witten and Frank, 2005) a collection of machine learning algorithms for data mining tasks, included in a Java API for easy integration.

The K-Means algorithm is a simple and easy procedure to classify a set of objects in a number of K groups (clusters), determining K a priori. It represents each of the groups by the average (or weighted average) of its points, that is, by its centroid, which is located in the center of the elements of each group. Representation by centroids has the advantage of having a graphic and statistical significance immediately. The Expectation-Maximization algorithm (EM) is an iterative method of maximum likelihood estimation.

To decide which of both algorithms to implement in our project we have focused on the time spent in its execution, since the results of the cluster were very similar between them.

We measure the time it takes to stage grouping, using the files generated by our application after four queries. Simple K-Means algorithm proved being much faster than EM algorithm, between 3 and 6 times faster, depending on the size of the documents set.

Based on these findings, we chose to continue working with the algorithm Simple-K-Means (with K=4), the best suited to our objectives.

3.2 Visualization on mobile devices

Access to the Web from mobile devices has a large number of constraints, as could be: screen size, no multiple windows, navigation limited, no JavaScript and cookies, page types available, slow speed, unbalanced pages from automatic conversion to mobile format, size messages, cost of shipping by size, etc…

The World Wide Web Consortium (W3C) is an international organization that works to develop web standards with general agreement. One of its working groups, the Mobile Web Initiative is a committee working on the creation of catalogs of best practices for creating mobile-friendly content, the easy access to device descriptions, the creation of test benches interoperability of mobile browsers

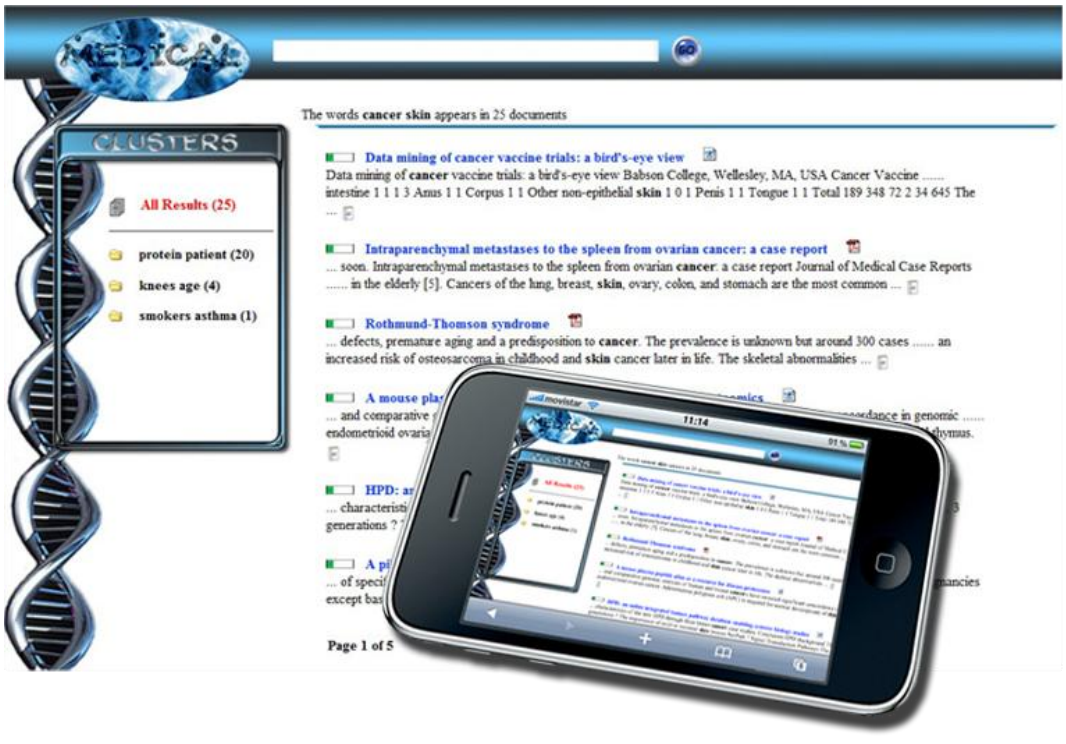

Fig.2. Viewing a cluster in high resolution format

and exploring ways to use the Web on mobile devices to reduce the digital divide.

(W3C, 2008) compiled a set of guidelines for designing and developing Web content for mobile devices. We've followed strictly the guide in the development of the tool presented in this work. As a result, we have implemented two user interfaces, one for devices with a larger screen (Fig. 2) (PC, notebook, Tablet PCs or UltraPCs) and one for smaller mobile devices (PDA's, Handhelds, Mobile, etc.).

4 Conclusions

The clustering of documents provides a mechanism for effective and intuitive navigation through the organization of the recovery results. When the system returns a large number of documents, organizing them in a small number of groups is very useful, especially on mobile devices with reduced screens.

In this paper we have presented a retrieval system for biomedical domain with post-retrieval clustering. Besides, we have presented two interfaces, one for desktop and one for mobile device.

Future efforts will focus on using a biomedical ontology, like UMLS Metathesaurus, to improve the clustering quality and the labeling of the groups.

Bibliography

Buenaga, M. de, Gachet, D., Maña, M., de la Villa, M., Mata J.:Clustering and Summarizing Medical Documents to Improve Mobile Retrieval. Workshop on Mobile Information Retrieval (MobIR'08) Singapore (2008)

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39 (1): 1–38 (1977)

Dunlavy, D.M., O'Leary, D.P., Conroy, J.M., Schlesinger, J.D. QCS: A system for querying, clustering and summarizing documents. Information Processing and Management 43(6), 1588--1605 (2007)

Garritty, 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), e7 (2006)

Gospodnetic, O., Hatcher E., McCandless M.: Lucene in Action (2nd ed.). Manning Publications (2009)

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; Nov-Dec; 14(6): 807-15 (2007)

Hearst, M., Pedersen, P.: Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. In: 19th Annual International ACM SIGIR Conference, pp. 76--84 (1996)

León, S.A., Fontelo, P., Green, L., Ackerman, M., Liu, F.: Evidence-based medicine among internal medicine residents in a community hospital program using smart phones. BMC Medical Informatics and Decision Making 7(5) (2007)

MacQueen, J. B.: Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297 (1966)

Maña, M.J., Buenaga, M., Gómez, J.M.: Multidocument summarization: An added value to clustering in interactive retrieval. ACM TOIS 22(2), 215--241 (2004)

Muñoz, M.A., Rodríguez, M., Favela, J., Martínez-García, A.I., González, V.M.: Context-aware mobile communication in hospitals. IEEE Computer 36(8), 38-46 (2003)

Rasmussen, E.: Clustering algorithms. En W. Frakes y R. Baeza-Yates, eds., Information Retrieval: Data Structures & Algorithms, pags. 419-442. Prentice-Hall International, London (1992)

Witten, I. H. y Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques (2nd ed.). Morgan Kaufmann, San Francisco. CA (2005)

W3C: Mobile Web Best Practices 1.0, Basic Guidelines. W3C Recommendation 29 July 2008. http://www.w3.org/TR/2008/REC-mobile-bp-20080729/ (2008)