# Bucefalo: A Tool for Intelligent Search and Filtering for Web-based Personal Health Records

Francisco P. Romero, Jesus Serrano-Guerrero, Jose A. Olivas,
SMILE Research Group
University of Castilla-La Mancha
Department of Information Technologies and Systems
Ciudad Real, Spain
{franciscop.romero, jesus.serrano, joseangel.olivas}@uclm.es

## ABSTRACT

In this poster, a tool named BUCEFALO is presented. This tool is specially designed to improve the information retrieval tasks in web-based Personal Health Records (PHR). This tool implements semantic and multilingual query expansion techniques and information filtering algorithms in order to help users find the most valuable information about a specific clinical case. The filtering model is based on fuzzy prototypes based filtering, data quality measures, user profiles and healthcare ontologies. The first experimental results illustrate the feasibility of this tool.

## Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval—*Information Filtering, Search Process*

## General Terms

Performance, Management.

## Keywords

Information Filtering, Web-based Personal Health Record.

## 1. INTRODUCTION

In the last years, with the expansion of the World Wide Web (WWW), the individually owned and controlled web-based personal health record (Google Health[1], Microsoft Health Vault[2] and NHS HealthSpace amongst others) have appeared. Therefore, the amount of accessible information about health care has increased enormously. This situation means that a vast amount of information of varying quality is disseminated. However, these new document repositories create new opportunities and challenges. In order to provide a more personalized and tailored service to their users, the need of an efficient and reliable information filtering process is critical.

The web-based PHRs use the health data exchange standards with the aim of representing clinical data. In these standards the relevant health information is reliably and unambiguously tagged using XML within a single file. The use of XML allows that this information can be read, understood and processed for any application which uses the standard. Google Health and Microsoft Health use a subset of the CCR (Continuity of Care Record) standard. The CCR standard is the most used patient health summary. A document in CCR format is a XML document that consists of a header, a footer, and a body of health data organized into as many as 17 sections, e.g. problems and conditions, medications list, allergies list, family history, procedures, encounters, etc..

These web-based PHRs are examples of multi-user document repositories. The clinical reports can be read for different users (nurses, physicians, students) and for different purposes (diagnosis, learning, research). When a document repository has many users and many purposes, there are different points of view of the same repository structure. Therefore, it is necessary a technique able to manage these different *points of view* in knowledge retrieval tasks. In this case, fuzzy logic is especially recommendable due to its special features to model information retrieval applications.

## 2. METHOD AND ARCHITECTURE

BUCEFALO provides a single interface for different sources of patient records. These web-based PHRs could be stored in Internet or in native XML databases (eXist, Software AG Tamino, etc.). The user only has to send a query and the system will answer the filtered and adapted results.

There are two main components in our tool, the search component and the filtering component (Fig. 1). The first increases the semantic capability of user queries and the second is used to organize the retrieved information. Both are important to improve the capability of retrieving information and to adapt the answer to each user. Therefore, both are based on user profiles and healthcare ontologies.

### 2.1 Search Step

The search step consists of the following phases in order to transform the original user query:

1. *Abbreviations Processing:* In medical records the use of abbreviations is very frequent. In order to process the query as fast as possible, the tool uses a simple efficient algorithm for extracting abbreviations and their definitions from biomedical texts implemented by Shwartz and Hearst [4].

[1] http://www.google.com/intl/es/health/about/index.html
[2] www.healthvault.com

*WWW 2009,* April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

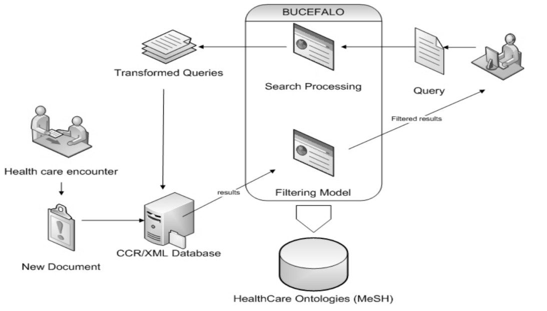

**Figure 1: Tool Architecture**

2. *Semantic Query Expansion:* Once the user sends a query to the system, the first step is to build a set of additional queries that complement the original one. In these queries new terms semantically related to those of the original one are included (synonyms, terms used to define the correct sense, concepts stored in ontologies like MeSH[3]). This process is guided through the user profile information because the relevant terms are not the same for different users groups.

3. *Multilingual Query Expansion:* Medical care is a multicultural and multilingual environment. This situation means that the information is disseminated in different languages, and therefore, it is more difficult to find the right information. When the tool processes non-English documents or queries, it is used a component called InterLingual Index (ILI) [2] to get the equivalent meanings between different languages and to expand the query using different languages.

## 2.2 Filtering Step

The large volume of results that are generated after the query execution is a major problem among users. Therefore it is necessary to filter and to retrieve relevant information. In this tool, the filtering model is based on four knowledge components with the aim of improving the precision of the obtained results.

- *Filtering based on Data Quality*: Applying data quality principles is a useful strategy for narrowing the search space in order to minimize computational costs and to increase the user satisfaction. In this tool some data quality filtering criteria such as the reliability, completeness and timeliness are used [1].

- *Filtering based on Fuzzy Prototypes* [3]: It is applied a special approach of category-based filtering. In this filtering method the documents are divided into categories using concepts that ocurr in each document. These categories are organized into a fuzzy hierarchical structure and represented as fuzzy deformable prototypes. The filtering process is performed using a conceptual matching among the fuzzy prototypes and the documents contents. Since filtering is a dynamic process, the hierarchy and the prototypes are automatically updated.

- *Fuzzy User Controller*: The user controller is a set of rules implemented by a fuzzy controller which represents the user preferences.

- *User Profile Filtering*: The tool uses a user profile with two components. The first component is an ontological definition of the professional categories (specialist, nurse, student), main search purposes (diagnosis, research, learning) and the user domain of knowledge. This representation is extracted from MeSH. The second component is the information feedback. Each obtained document has associated a relevance feedback button asking the user for scoring the usefulness of that result.

# 3. PRELIMINARY EXPERIMENTS

Some experiments have been carried out in order to analyze the tool perfomance. First, we have built a collection of CCR documents using clinical cases extracted from Internet and CCR templates. This collection has been stored in a native XML database. Next, a test experiment has been designed with three health care proffesionals: a nurse, a physician and a researcher. In the experiment, they have analyzed ten queries transformed and expanded according to their user profiles. In this test we achieve an 85% of correct transformations. The second user test consisted in analyzing the results obtained after the filtering process using the above-mentioned queries. In this test the average number of relevant documents that were filtered was the 80%.

# 4. CONCLUSIONS AND OPEN ISSUES

This work presents a tool with the aim of improving the retrieval tasks in web-based Personal Healtcare Records based on CCR XML documents. This approach will provide more relevant documents for the users, due to the consideration of their user profiles in the making search decisions and the use of several techniques of query expansion and category-based filtering. At this moment, this tool is only a prototype. The next step is assessing BUCEFALO with real world sources of CCR documents and taking in account users' dynamically changing criteria of relevance. Nevertheless, several tests with particular collections of documents have been carried out, obtaining good results.

# 5. REFERENCES

[1] I. Caballero and E. Verbo. A Data Quality Measurement Information Model based on ISO/IEC 15939. In *Proceedings of the 12th International Conference on Information Quality*, 2007.

[2] J. Ellman. Eurowordnet: A multilingual database with lexical semantic networks. *Nat. Lang. Eng.*, 9(4):427–430, 2003.

[3] F. P. Romero, J. A. Olivas, and P. J. Garces. Inference based on fuzzy deformable prototypes for information filtering in dynamic web repositories. In *Proceedings of the 2007 IEEE International Conference on Fuzzy Systems*, pages 1403–1408, 2007.

[4] A. Schwartz and M. Hearst. A simple algorithm for identifying abbreviation definitions in biomedical texts. In *Proceedings of the Pacific Symposium on Biocomputing (PSB 2003)*, pages 451-462, 2003.

[3] http://www.nlm.nih.gov/pubs/factsheets/mesh.html