Biomedical Literature Retrieval Based on Patient Information

Ana Jimenez-Castellanos[1], Izaskun Fernandez[2], David Perez-Rey[1], Elisa Viejo[3], Francisco Javier Díez[2], Xabier García de Kortazar[2], Miguel Garcia-Remesal[1], Victor Maojo[1], Antonio Cobo[4,3], and Francisco del Pozo[3,4]

[1] Dept Inteligencia Artificial, Facultad de Informática
Universidad Politécnica de Madrid, Campus de Montegancedo S/N
28660 Boadilla del Monte, Madrid, Spain
[2] Tekniker-IK4, Av Otaola, 20, 20600 Eibar, Guipuzcoa, Spain
[3] Centre for Biomedical Technology, Technical University of Madrid
Campus de Montegancedo - UPM Autopista M-40, km 38. Pozuelo de Alarcón
28223 Madrid, Spain
[4] Biomedical Research Networking Center in Bioengineering
Biomaterials and Nanomedicine, Madrid, Spain
{ajimenez,dperez,mgarcia,vmaojo}@infomed.dia.fi.upm.es,
{ifernandez,fdiez,xkortazar}@tekniker.es,
{evdiego,acobo,fpozo}@gbt.tfo.upm.es

Abstract. Information and Communication Technologies has led to a biomedical data explosion. A proportional growth has been produced regarding the amount of scientific literature, but information retrieval methods did not follow the same pattern. By using specialized clinical search engines such as PubMed, Medscape and Cochrane, biomedical publications has became instantly available for clinical users. However, additional parameters, such as user context, are not taken into account yet. Initial queries still retrieve too many results without a relevance-based ranking. The objective of this work was to develop a new method to enhance scientific literature searches from various sources, by including patient information in the retrieval process. Two pathologies have been used to test the proposed method: diabetes and arterial hypertension. Results obtained suggest the suitability of the approach, highlighting the publications related to patient characteristics.

Keywords: Electronic health record, Search engines, Literature retrieval, Integration, Federated search.

1 Introduction

Physicians attending daily a high number of patients hamper their availability to keep up to date with the latest research news in their field. They frequently lack the time to locate relevant information related to the patient Electronic Health Record (EHR). Physicians need specific information rather than large amounts of information, however the amount of data generated nowadays and stored both, within EHR and research literature, is overwhelming. The more data the more challenging is to find relevant information (for

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 312–323, 2013.

research and training of the physician) and to integrate it with current resources. Search engines and other Web 2.0 technologies such as RSS feeds facilitate these tasks, but advanced methods are required to retrieve research information related to, not only a pathology, but also a specific demographic group.

Scientific research papers regarding clinical practice deals with population groups rather than specific patients. Physicians must generalize patient data in order to find patient-related clinical literature. They should analyse the EHR of the patient to include relevant keywords in the query that filter the results from several information sources. After that, the physician reviews the scientific articles retrieved and selects the relevant information for the patient, discarding the majority of the results. This is particularly interesting for patients who do not respond to therapies or treatments, who need special attention to avoid further complications. These kind of patients are those for whom physicians need more specific research information about their disease.

In this work we present a new method, implemented by the authors, aiming to provide an advanced method to retrieve biomedical literature based on EHRs. Diabetes and arterial hypertension are the two use cases that have been tested, within the framework of *Tratamiento 2.0*, a research and development project aiming to create a generic middle ware platform that serves as the basis for the development of management services and intelligent application for treatment for patients, especially chronic [1].

2 Background

Patient information inclusion in search queries for biomedical literature, requires techniques and technologies with a high research activity. Those used to store the required information, such as Electronic Health Records and Decision Support Systems, and those used to extract information, such as Natural Language Processing and Federated Searching described in this section.

Electronic Health Records (EHR), defined as electronic objects that contain data, evaluations and information of any kind, on the status and the clinical course of one patient through the care process [2], have been a fundamental advance in clinical practice. Electronic information can be automatically analysed and there exist research efforts to generate knowledge from EHRs such as [3], where the data generated from care delivery and captured in the EHR systems, is used for being analysed to discover and then inform about best practice. In [4], how to transfer knowledge from a medical record written in a free text form into a structured format represented by the EHR is analysed. Finally, [5] is a study aiming to understand user needs as captured by their search queries in an EHR system.

EHR are the essential infrastructure to other software elements in clinical practice such as Decision Support Systems, a kind of information systems that supports decision-making activities [6]. Specifically, Clinical Decision Support Systems (CDSS) are software programs that assist the physicians in the decision making process [7][8]. Some methods to integrate CDSS and EHR are: [9] and [10], systems dealing with CDSS adaptation, to facilitate the integration of individual patients preferences and characteristics and improve decision making.

To retrieve further clinical information, the EBM [11] recommends physicians to formulate clinical questions in terms of the problem/population, intervention, comparison, and outcome for searching clinical reports efficiently. These elements comprise the PICO frame: P represents problem/population; I interventions information; C the comparison; and finally O the outcome. The construction of this kind of query is not an easy task as it is shown in [12]. For instance, defining P requires an exhaustive reading of the patient EHR and the selection of the most relevant characteristics for the current context.

Information extraction dealing with free text instead of structured data is treated with Natural Language Processing (NLP) techniques, that have been applied to many tasks of Biomedicine such as bio-entity recognition, protein/gene normalization, interaction, extraction and many others. As there are many fields where NLP techniques can be used, various NLP tools are available. There are domain specific tools resolving an specific task like GoAnnotatorTool [13], but also we can find more generic and flexible NLP tools, like Freeling [14] or GATE [15] which can be used for many tasks and domains. The fundamental problem of NLP analysis is the fact, that a particular meaning may be expressed using different synonymous expressions. Lexical repositories recording meaning and lexical forms relations are frequently used in NLP techniques for interpreting texts. MeSH [16] or SNOMED [17] are two well known lexical repositories in the medicine domain which can be exploited for scientific article interpretation.

Finally, heterogeneous data sources stored in different locations require distributed or federated methods to extract information. The federated search paradigm was thus created, evolving in response to the vast number of information resources. As defined in [18], federated searching consists of firstly transforming a query and broadcasting it to a group of different sources; merging the results obtained from the different asked sources; presenting it in a unified format without duplications; and providing the option of sorting the result set. Search engines like Sphinx [19] help in the development of this kind of systems, indexing the information and providing a fast information retrieval.

These techniques have been used in projects such as Parallel IE for bio-medical text mining and indexation at Merck kGaA, Darmstadt and Medline Analysis [20] at Institute for Medical Informatics and Biometry, University of Rostock, Germany for extracting causal functional relations on MedLine abstracts using GATE. There are also tools like MeSHMap [21] and MedMeSH Summarizer [22] that exploits MeSH ontology for document indexation and summarization, respectively.

3 Patient-Based Literature Retrieval and Integration

The approach proposed in this paper aims to provide an improved method to search biomedical literature based on patient data. Data used to identify the relevant information for literature retrieval is collected from two sources: sensors on the patients home described elsewhere [1] and data contained in the EHR of the patient. Both sources are stored in the platform of the project and then feed the services explained below.

Fig. 1 presents the two main phases performed to retrieve relevant publications. Firstly, the system should receive both data sources and locate relevant characteristics of the patient and then looking for and integrating the most relevant publications based on these characteristics.

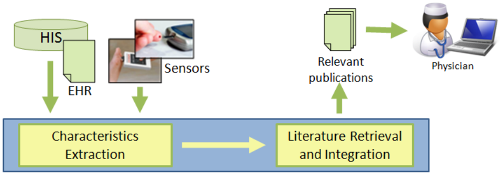


Fig. 1. Patient-based literature retrieval architecture

To automatically query current biomedical literature search engines for relevant publications based on EHRs and sensor info, a web services architecture was proposed. Three main web services were identified to cover the required functionality:

– *patientCharacterictics*, receiving a set of parameters from the EHR and sensors to identify the relevant characteristics.
– *federatedSearch*, receiving the relevant characteristics, generates and launches queries against different data sources.
– *resultAggregation*, that collect, integrates and present the results to the user.

These web services are explained within the next subsections, 3.1, 3.2 and 3.3 respectively.

3.1 Characteristics Extraction

The *patientCharacteristics* web service receives a set of parameters in XML format from the platform of the project. This platform stores data contained in the patient's EHR in addition to other data resulting from the monitoring of the patient at home e.g. certain data is needed in real time (like glucose levels before and after eating). This was implemented using sensors within the patient environment. Fig. 2 shows the structure of the web service which extracts the characteristics of a patient.

Two separated phases can be identified: (i) parameter processing and (ii) curation. Within the first phase, parameters are processed to extract the relevant characteristics. Afterwards, during the curation phase, an expert may evaluate the characteristics extracted to filter those relevant for the publication search.

Parameters from EHR or sensors are integrated into a single XML parameter list:

```
<patient_parameters>
   <pathology>diabetes</pathology>
   <parameters_list>
      <parameter>
         <name>age</name>
         <value>62</value>
      </parameter>
      ...
   </parameters_list>
</patient_parameters>
```

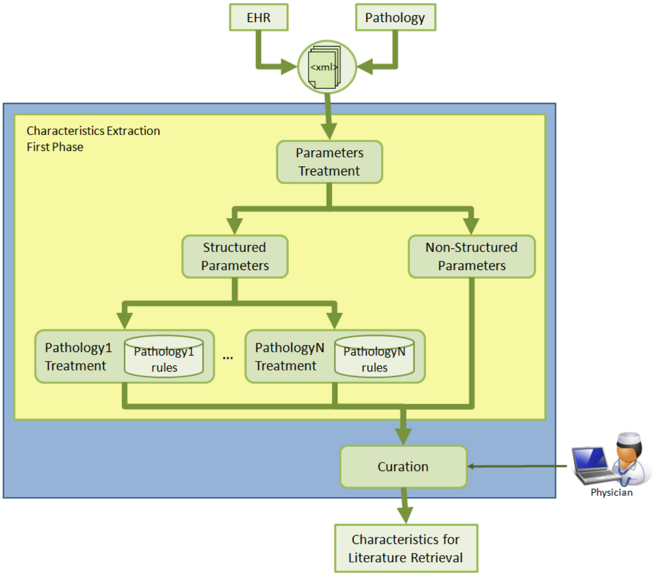


Fig. 2. Characteristics Extraction

For each parameter its name and its value are required. The *patientCharacteristics* web service extracts the significant characteristics focused on a given pathology.

The set of parameters is not closed, the service can receive any type of parameters, i.e. that the application will analyze only the parameters that can be treated according to certain rules, discarding those that are not recognized.

After receiving the parameters, the web service separates among structured and non-structured parameters. Structured data are those which have a numeric or enumerated value (E.g.: the numeric value of the patient's age, or the values *male* or *female* identifying the sex of the patient). Not structured data are free text and they are directly addressed to the curation phase as characteristics of the patient.

Depending on the pathology, structured parameters are selected for its relevance. For each disease there is a knowledge base of rules that is responsible for processing related parameters. Thus, when the characteristics of a patient have to be extracted based on a new disease, only the corresponding rules are required. The knowledge base of each pathology is built based on standards of CDSSs.

Relevant data for the pathology is analysed following rules contained within the knowledge bases, which are associated with the parameters. Those rules deal with patient stratification. Most of the rules are constructed based on expert approved thresholds. Depending on parameters values, characteristics are extracted and mapped to one or several MeSH terms.

All parameters are evaluated to extract significant characteristics related to one patient. In the curation phase, the physician (using a simple interface) may choose characteristics to be used to build the search query. Relevant characteristics selected by the physicians will be sent to the next web service for retrieving the corresponding literature.

3.2 Query Generation

Search federation intends to aggregate information units from heterogeneous sources, presenting the results on a unified way, without duplicates and ordered by relevance. This architecture enables a high scalability and stability, since adding new sources is only a matter of new connectors configuration, and system remains working even if some sources are down.

So the *federatedSearch* web service gets as input the XML characteristics generated by *patientCharacteristics* service as context and a query defined by the expert, which expresses the searching topic. In order to represent efficiently those characteristics and the free query, the service exploits MeSH ontology combined with NLP tools, trying to find corresponding MeSH terms both, for the characteristics and on the input free text. The MeSH lexicon as a flexible Gazetteer on GATE tool has been integrated in the service. In this way MeSH terms on any text are tagged, first parsing it with the morphological module of GATE and then identifying MeSH terms with the gazetteers module.

The characteristics are extracted from the patient EHR, so it can be considered that the selected MeSH terms define the problem (P) of the PICO query. Intervention (I) and outcome (O) will be completed with the identified and after selected terms on the expert query.

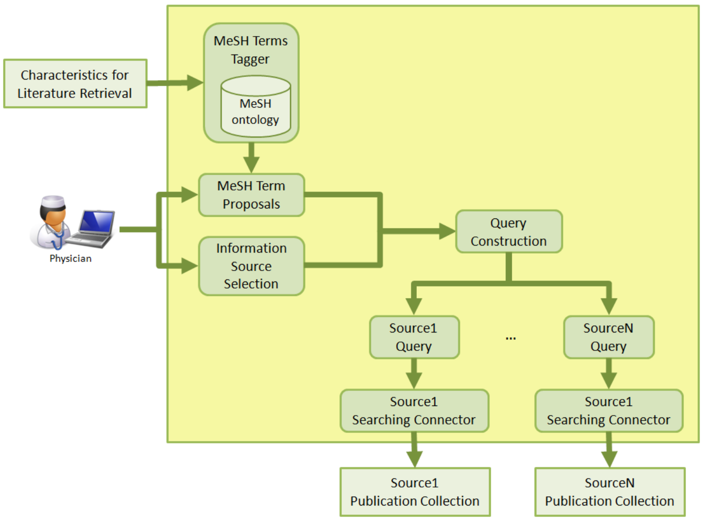


Fig. 3. Query Generation and Information Retrieval

In order to get the federated search, the web service queries different information sources. Each registered expert has associated a collection of information sources. For each new query the expert selects a set of those sources to search. Each information source encloses its own query format. Before searching, it is necessary to transform the

PICO query to the each source query format. For the transformation process, regardless of the source, the system applies always the same methodology: taking PICO query as input, it applies an xslt based transformation. The xslt object is a source dependent element where the source specific query syntax is considered for the final query representation. Together with query representations, the service launches different queries, getting one result set from each source in XML format. This procedure is graphically represented at Fig. 3.

3.3 Integration and Presentation

The publication sets obtained from the previous web service are the input of the *result-Aggregation* service. The aim of this service is to aggregate all the result sets presenting to the expert the publications as a unique collection. The final collection should not have duplicated items and must be sorted by relevance regardless of the source. The web service processes the publications, tagging MeSH terms and storing both the item itself and the tagged MeSH terms in a MySQL database in order to index the entire collection with Sphinx and improve the efficiency of the sorting search.

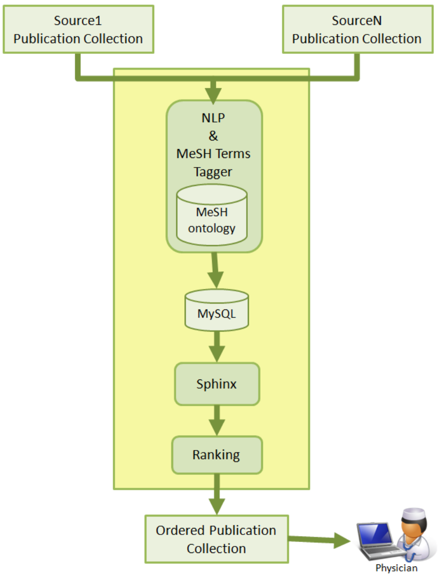


Fig. 4. Information Aggregation

Once the collection is indexed, the service ranks the publications based on the initial query using Sphinx. The result is a sorted collection based on the relevance of each publication respect the initial query as it is shown in Fig. 4.

4 Use Cases: Diabetes and Arterial Hypertension

Two use cases were developed to test the proposed method: (i) diabetes and (ii) arterial hypertension. A knowledge base to analyze the parameters related with them has been implemented to test the system. Information needed to evaluate each parameter received by the *patientCharacteristics* web service is stored in different knowledge (one for each pathology). Structured parameters were selected by experts within the project using a focus group methodology. These parameters are shown in table. 1.

Table 1. Recognized structured data

Common parameters	Diabetes specific	Arterial Hypertension
age	diabetic_familiar	hypertensive_familiar
race	neuropathy	total_cholesterol
sex	basal_glycemia	triglycerides
body_mass_index	gestational_diabetes	systolic_pressure
	retinopathy	diastolic_pressure
	nephropathy	HDL
	cardiopathy	abdominal_wall_size
	glucose_intolerance	smoker
	glycated_hemoglobin	cLDL
	diabetic_foot	

Depending on the pathology, structured parameters are selected by its relevance. Due to the characteristics of the diseases, there are some parameters that are common to both and others are specific for each pathology (table. 1).

Finally, for each structured parameter, the Web Service applies several rules, extracting the related characteristic. For example, treatment of the glycated hemoglobin (HbA1c) and related characteristics are:

- HbA1c $<$ 6.5 : "normal HbA1c"
- HbA1c $<$ 7 : "high HbA1c"
- HbA1c $<$ 8 : "high HbA1c AND glycemic control"
- HbA1c $>=$ 8 : "high HbA1c AND intensive glycemic control"

Once all the characteristics are extracted went through the curation phase, the system constructs a searching query with the pathology, the relevant characteristics previously gathered and the MeSH terms identified with the experts input query (with the searching topic).

With queries close to the mentioned pathologies, the system searches on expert defined sources using their corresponding connectors to obtain the most relevant publications and evidences. Nowadays, sources covered by the system are PubMed, Cochrane and MedScape, where the most relevant publications about diabetes and hypertension are published.

Finally, the results of each source are integrated, indexed and presented to the expert as a unique collection as shown in the following section.

5 Results

Characteristics are extracted according to the pathology and the parameters introduced. From this output, the expert will select the most relevant features to be launched to the search.

A preliminary test set of ten patient data and fifty queries were used to check the functionality of the system. The retrieved characteristics were correct for all the cases. The results correctly identified the 90% of the relevant papers, according to assessment of experts of the project.

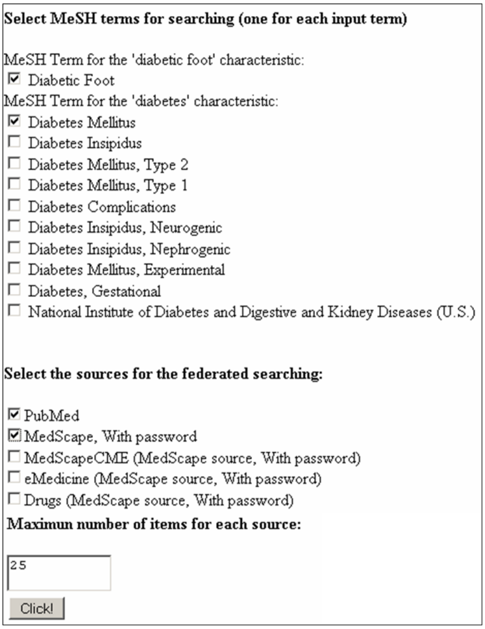
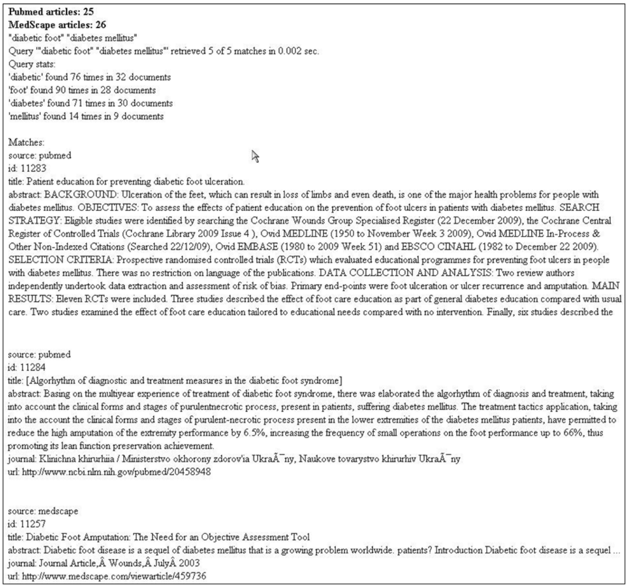


Fig. 5. Interface to select the parameters and sources of the search

As it is shown in Fig. 5, the expert can choose a more specific term for the search, i.e. optimizing the query that will be launched against the selected sources. The expert can also specify the number of results obtained from the query.

With the optimized query, the system accesses the different sources and aggregates the results, filtering those fulfilling the initial requirements and finally presenting them ordered by relevance, as it is shown on Fig. 6.

Two filters were applied to the parameters in order to extract the most relevant literature referred to a patient: (i) a transformation of the parameters into Mesh terms and (ii) an optimization of the characteristics, using the MeSH ontology and the medical knowledge of the expert. Through the final query and the federated search the expert will retrieve biomedical literature based on patient data.


Pubmed articles: 25
MedScape articles: 26
"diabetic foot" "diabetes mellitus"
Query '"diabetic foot" "diabetes mellitus"' retrieved 5 of 5 matches in 0.002 sec.
Query stats:
'diabetic' found 76 times in 32 documents
'foot' found 90 times in 28 documents
'diabetes' found 71 times in 30 documents
'mellitus' found 14 times in 9 documents

Matches:
source: pubmed
id: 11283
title: Patient education for preventing diabetic foot ulceration.
abstract: BACKGROUND: Ulceration of the feet, which can result in loss of limbs and even death, is one of the major health problems for people with diabetes mellitus. OBJECTIVES: To assess the effects of patient education on the prevention of foot ulcers in patients with diabetes mellitus. SEARCH STRATEGY: Eligible studies were identified by searching the Cochrane Wounds Group Specialised Register (22 December 2009), the Cochrane Central Register of Controlled Trials (Cochrane Library 2009 Issue 4), Ovid MEDLINE (1950 to November Week 3 2009), Ovid MEDLINE In-Process & Other Non-Indexed Citations (Searched 22/12/09), Ovid EMBASE (1980 to 2009 Week 51) and EBSCO CINAHL (1982 to December 22 2009). SELECTION CRITERIA: Prospective randomised controlled trials (RCTs) which evaluated educational programmes for preventing foot ulcers in people with diabetes mellitus. There was no restriction on language of the publications. DATA COLLECTION AND ANALYSIS: Two review authors independently undertook data extraction and assessment of risk of bias. Primary end-points were foot ulceration or ulcer recurrence and amputation. MAIN RESULTS: Eleven RCTs were included. Three studies described the effect of foot care education as part of general diabetes education compared with usual care. Two studies examined the effect of foot care education tailored to educational needs compared with no intervention. Finally, six studies described the

source: pubmed
id: 11284
title: [Algorhythm of diagnostic and treatment measures in the diabetic foot syndrome]
abstract: Basing on the multiyear experience of treatment of diabetic foot syndrome, there was elaborated the algorhythm of diagnosis and treatment, taking into account the clinical forms and stages of purulentnecrotic process, present in patients, suffering diabetes mellitus. The treatment tactics application, taking into the account the clinical forms and stages of purulent-necrotic process present in the lower extremities of the diabetes mellitus patients, have permitted to reduce the high amputation of the extremity performance by 6.5%, increasing the frequency of small operations on the foot performance up to 66%, thus promoting its lean function preservation achievement.
journal: Klinichna khirurhiia / Ministerstvo okhorony zdorov'ia UkraÃ¯ny, Naukove tovarystvo khirurhiv UkraÃ¯ny
url: http://www.ncbi.nlm.nih.gov/pubmed/20458948

source: medscape
id: 11257
title: Diabetic Foot Amputation: The Need for an Objective Assessment Tool
abstract: Diabetic foot disease is a sequel of diabetes mellitus that is a growing problem worldwide. patients? Introduction Diabetic foot disease is a sequel ...
journal: Journal Article,Â Wounds,Â JulyÂ 2003
url: http://www.medscape.com/viewarticle/459736

Fig. 6. Aggregated Results

6 Conclusions and Future Lines

In this work, a new method for search clinical literature has been proposed and implemented within the framework of a research project for the personalization of treatments and therapeutic strategies. Promising results of the system suggest that EHR-based literature retrieval may facilitate the work of physicians obtaining biomedical patient related research publications. Since there is not any gold-standard corpus defined for evaluating the system empirically, we have done a qualitative evaluation of it testing with real patients and experts from the Hospital Universitario de Valencia and most of the clinicians were satisfied with the usability of the system for their daily practice.

As work in progress, extensions of services described within this paper are being implemented: further non-structured data processing, clustering techniques to group the results, open of generic connectors and source weighting for final ranking.

First, the service will use Natural Language Processing techniques to process non-structured data of the patient for characteristic extraction.

Clustering techniques are being used to present results not only as a sorted collection but also grouped by their semantic similarity, providing a more intuitive representation for the navigation.

Regarding information sources, connector designing process is being generalized to facilitate new source integration within the system. This service will add a weighting functionality for sources, to adjust the final publication relevance not only based on the initial query but also considering the relevance of the source.

The system presented in this work has proven the suitability of a patient-based literature retrieval approach. And we are confident that current extensions will improve the location of relevant publications for physicians, facilitating more relevant results when searching at main biomedical literature search engines.

Acknowledgements. The present work has been funded by the Spanish Ministry of Industry, Tourism and Trade in the framework of the ”Avanza I+D” programme under the project *Tratamiento 2.0* (project number TSI-020301-2008-15). We would like to express our gratitude to the *Hospital Universitario de Valencia* for their collaboration in the election of the parameters used in this work.

References

1. Tratamiento 2.0, `http://www.tratamiento20.com`
2. Healthcare Information and Management Systems Society: Consensus definition of an Electronic Health Record, `http://www.himss.org/ASP/topics_ehr.asp`
3. Weaver, C.A., Warren, J.J., Delaney, C., et al.: Bedside, Classroom and Bench: Collaborative Strategies to Generate Evidence-Based Knowledge for Nursing Practice. Int. J. Med. Inform. 74, 989–999 (2005)
4. Antolik, J.: Automatic Annotation of Medical Records. Stud. Health Technol. Inform. 116, 817–822 (2005)
5. Natarajan, K., Stein, D., Jain, S., et al.: An Analysis of Clinical Queries in an Electronic Health Record Search Utility. Int. J. Med. Inform. (2010)
6. Keen, P.G.W., Morton, M.S.S.: Decision support systems: An organizational perspective. Addison Wesley Publishing Company (1978)
7. Kawamoto, K., Houlihan, C.A., Balas, E.A., et al.: Improving Clinical Practice using Clinical Decision Support Systems: A Systematic Review of Trials to Identify Features Critical to Success. BMJ 330, 765 (2005)
8. Sim, I., Gorman, P., Greenes, R.A., et al.: Clinical Decision Support Systems for the Practice of Evidence-Based Medicine. J. Am. Med. Inform. Assoc. 8, 527–534 (2001)
9. van der Weijden, T., Legare, F., Boivin, A., et al.: How to Integrate Individual Patient Values and Preferences in Clinical Practice Guidelines? A Research Protocol. Implement Sci. 5, 10 (2010)
10. Seidling, H.M., Schmitt, S.P., Bruckner, T., et al.: Patient-Specific Electronic Decision Support Reduces Prescription of Excessive Doses. Qual. Saf. Health. Care (2010)
11. Evidence-Based Medicine Working Group: Evidence-Based Medicine. A New Approach to Teaching the Practice of Medicine. JAMA 268, 2420–2425 (1992)
12. Huang, X., Lin, J., Demner-Fushman, D.: Evaluation of PICO as a Knowledge Representation for Clinical Questions. In: AMIA. Annu. Symp. Proc., pp. 359–363 (2006)
13. Couto, F.M., Silva, M.J., Lee, V., et al.: GOAnnotator: Linking Protein GO Annotations to Evidence Text. J. Biomed. Discov. Collab. 1, 19 (2006)
14. Atserias, J., Casas, B., Comelles, E., et al.: FreeLing 1.3: Syntactic and Semantic Services in an Open-Source NLP Library. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), pp. 48–55 (2006)
15. Cunningham, H.: GATE, a General Architecture for Text Engineering. Computers and the Humanities 36, 223–254 (2002)
16. Díaz-Galiano, M.C., García-Cumbreras, M.Á., Martín-Valdivia, M.T., Montejo-Ráez, A., Ureña-López, L.A.: Integrating MeSH Ontology to Improve Medical Information Retrieval. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 601–606. Springer, Heidelberg (2008)

17. Spackman, K.A., Reynoso, G.: Examining SNOMED from the Perspective of Formal Ontological Principles: Some Preliminary Analysis and Observations. KR-MED., 72–80 (2004)
18. Jacso, P.: Thoughts about Federated Searching. Information Today 21, 17 (2004)
19. Lee, K.F.: Automatic speech recognition: The development of the SPHINX system. Kluwer Academic Pub. (1989)
20. Karopka, T., Scheel, T., Bansemer, S., et al.: Automatic Construction of Gene Relation Networks using Text Mining and Gene Expression Data. Med. Inform. Internet Med. 29, 169–183 (2004)
21. Srinivasan, P.: MeSHmap: A Text Mining Tool for MEDLINE. In: Proc. AMIA. Symp., pp. 642–646 (2001)
22. Kankar, P., Adak, S., Sarkar, A., et al.: MedMeSH Summarizer: Text Mining for Gene Clusters, pp. 11–13 (2002)