A New Method to Retrieve, Cluster And Annotate Clinical Literature Related To Electronic Health Records

Izaskun Fernandez[1], Ana Jimenez-Castellanos[2], Xabier García de Kortazar[1], and David Perez-Rey[2]

[1] Tekniker-IK4
{ifernandez,xkortazar}@tekniker.es
[2] Dept. Int. Artificial, Facultad de Informática, Universidad Politécnica de Madrid
{ajimenez,dperez}@infomed.dia.fi.upm.es

Abstract. The access to medical literature collections such as *PubMed*, *MedScape* or *Cochrane* has been increased notably in the last years by the web-based tools that provide instant access to the information. However, more sophisticated methodologies are needed to exploit efficiently all that information. The lack of advanced search methods in clinical domain produce that even using well-defined questions for a particular disease, clinicians receive too many results. Since no information analysis is applied afterwards, some relevant results which are not presented in the top of the resultant collection could be ignored by the expert causing an important loose of information. In this work we present a new method to improve scientific article search using patient information for query generation. Using federated search strategy, it is able to simultaneously search in different resources and present a unique relevant literature collection. And applying NLP techniques it presents semantically similar publications together, facilitating the identification of relevant information to clinicians. This method aims to be the foundation of a collaborative environment for sharing clinical knowledge related to patients and scientific publications.

Keywords: Electronic Health Record, search engines, literature retrieval, integration, federated search, collaborative environment

1 Introduction

Web technologies have produced an explosion in the production and availability of clinical publications. Access to massive amounts of information by the practitioners, has led to include literature search tools to support patient diagnosis and treatment. However, clinical publications are focused on population groups rather than specific patients. So in order to find relevant publications that fit with specific characteristics of a particular patient, clinicians must generalize the characteristics of the Electronic Health Record (EHR) to meet the generic patterns of the relevant literature.

Clinicians accessing PubMed , MedScape or Cochrane for searching publications related to patients frequently get too many results which they should revise in order to recover relevant information. In addition, biomedical information is nowadays distributed across different repositories, so they have to execute multiple queries in the distributed resources to retrieve publications regarding a specific patient and pathology. In other areas, meta-search engines and other tools have been widely used facilitating the integration of information from multiple sources. However, in clinical practice, there are few trials exploiting the potential of such cutting-edge technologies.

In this work we propose a method aiming to provide a robust environment for searching literature in the daily clinical practice of the physicians. The proposed method is based on previous works regarding EHR-based literature retrieval [1], improving the integration of searching results and extending its functionality with publications clustering and annotation strategies.

The article is structured as follows: the next section presents related works; in section 3 the method for retrieving and analysing biomedical literature based on EHRs and clusterization techniques is described; section 4 presents the obtained results and evaluation within the *Tratamiento 2.0* project environment; to finish, conclusions and future works are explained.

2 Background

In recent years there has been an increasing interest in extracting and enriching the content of EHR standards ([2], [3]) such as: Health Level 7 Clinical Document Architecture (HL7-CDA) [4] and openEHR [5]. Since the majority of the content of EHR and clinical data is free text instead of structured information, Natural Language Processing techniques are frequently required. For instance, the method presented in [7] is based on GATE (General Architecture for Text Engineering) and the UMLS standard vocabulary to extract specific information from pathology reports.

Previous works also exist on discovering biomedical relationships from semantic annotations [8] [9], but they are limited to present just PubMed abstract collections as result. In [1], the authors presented a system to support clinicians in the literature search process using EHR information. For literature retrieval, the federated search methodology [10] was used for integrating results obtained from different information sources, pre-selected by the user. Search engines like Sphinx [6] help in the development of this kind of systems, indexing the information and providing a fast information retrieval.

In information retrieval systems, the query formulation is a crucial step to obtain suitable results. The EBM (Evidence-Based Medicine Working Group 1992) recommends to create clinical questions using the PICO frame. Improved results were obtained using the PICO query format [11] [12] but, the creation of this queries is not a trivial task as it is shown in [13].

The literature in the area describes numerous efforts on both, extracting information from EHR and improving accuracy and efficiency of clinical searches, but new tools are required to provide support to clinicians.

3 Method

We propose a method with a modular architecture to address the requirements of an EHR-based literature process. The main modules are: a query generation module, a federated-search module, an automatic literature processing process and finally a publication annotation module. The architecture is graphically presented at Fig.1

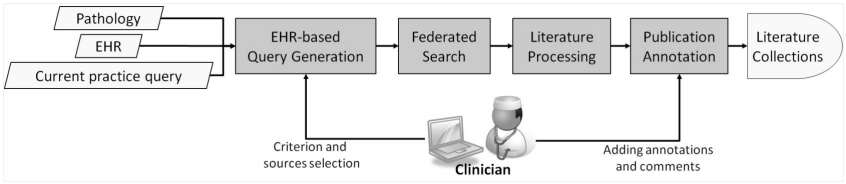

Fig. 1. Method's architecture

The *EHR-Based Query Generation* module deals with electronic health record information, extracting the most relevant features regarding specific pathologies. These characteristics are used to give context to the searching criteria specified by the expert for construct the final query, which will be used in the searching phase to query different repositories in the second module. The method manages every independent result set obtained from each repository, presenting the result as a unique and integral literature collection, semantically annotated and clustered in the third module. Within the fourth module, clinicians may also annotate and comment the publications of the collection.

3.1 EHR-Based Query Generation

The aim of the first module is to support the clinician on literature query generation for a certain criterion regarding an individual patient and disease, formulating the query based on PICO model. The method combines both the relevant information of the individual patient for the particular disease and the theme about the clinician wants to search about. A PICO query is composed by four main elements, P,I,C and O: P represents the problem/population; I intervention's information; C the comparison; and finally O the outcome. When constructing PICO queries not always all the elements are necessary.

Let's assume that a clinician wants to search about if there is any contraindication to prescribe *ECA inhibitors* (described as current criterion) for *Patient A* that is a *chronic diabetic* patient (previously mentioned as pathology). In PICO

format, the P element should be represented by the most relevant characteristics of Patient A about his/her chronic diabetes; the I element by *ECA inhibitors* drug; and the O element by *contraindication*, which is the outcome the clinician wants to know about. For this query, since no comparison criterion is used, the C element would be empty.

Since the aim of this module is to automatize the process of extracting and assigning values to PICO components, it applies different strategies to the input parameters. For P definition, the module uses a a web service that extracts the relevant characteristics from the input EHR about a certain pathology. As it is detailed in [1], the web service uses a knowledge base (KB) which defines the relevant parameters and their values. And applying the transformation rules in the KB it translated the structured or numeric values into concepts for P, such as *high glucose*.

For completing I, C and O components in a efficient way, minimizing the non relevant information expressed on the clinician's expressions concerning to the current practice to search about, the module applies a NLP process. A NLP process which takes as input a Spanish or English free text and extracts the most relevant terminology. For that purpose it uses a GATE[3] application which combines a tokenizer, sentence splitter, part of speech tagger, stemmer and finally an ontology based gazetteer tagger, and tags all the expressions in the text referring to any term in the ontology. The stemmer application is a crucial step in order to identify not only the perfect matching expressions but also the inflected mentions of the ontology vocabulary, such as *potassium ion levels*, the pluralized form of *potassium ion level* ontology concept. The ontology for the gazetteer tagger is parametrizable, which makes the module portable for different pathologies and domains.

The *EHR-Based Query Generation* interacts with the clinician, presenting the automatically extracted terminology and he/she must select the most relevant ones for each PICO field.

3.2 Federated Search

Federated search module allows searching different repositories simultaneously with the same query. In this module clinicians can not only select the desired resources from a predefined list, but also they can weight the selected resources establishing preferences among them. The influence of these preferences in results will be sown at 3.3 section.

The query defined in the previous module is used to search every selected source. Since each resource should manage the information in a different way, a configuration file is associated to each source defining the way to ask and how to get results from it. Concretely, the configuration file specifies how the PICO query should be transformed for asking the current resource and also, how to

[3] General Architecture for Text Engineering is a open source software capable of solving almost any text processing problem. More information at `http://gate.ac.uk/`

interpret the searching output. The configurations files are implemented using *xslt*, a declarative language designed for defining XML file transformations.

Using *xslt* files for specifying source dependent characteristic instead of including them in the module code, makes it flexible. Flexible in the sense that defining a *xslt* configuration file is enough for including a new source in the module.

So the module accesses to the configuration files of the selected sources transforming the PICO query in the corresponding formats, asks the sources and again, using the configuration files, it interprets each output and stores and merges all the publications result sets in the server database as a unique collection. For each publication this module gets the title (required), and a short description, authors and the publication date when they are available. Since some sources can share literature, the method eliminates duplicated publications using title information.

3.3 Literature Processing

Literature collections without duplications are used in the following steps to present not only a list of relevant publications, but also an enriched set of scientific literature. In the server database for each publication among others, there is available at least the title and a short description if it is extracted at searching phase. Applying to that literature content the NLP terminology extractor process described at Section 3.1, this module obtains the relevant terminology on each publication. Since each publication is annotated separately, the method can access to the terminology of each item on the collection and present the entire collection as: (i) a content based relevance ranked list, (ii) a list of publications ordered by both content and source relevance or (iii) a semantically clustered collection.

In the first two options, using *Sphinx index engine*[4] the method compares the extracted literature terminology with the query terminology. It ranks the publications measuring the shared terminology in both, query and publication, and it adjusts the measure with the resource weight for the (ii) modality. This way in (ii) option publications are ordered by their the relevance respect to the query, but also by the clinician resource preferences using the belonging resource weights defined in the federated search module (described at 3.2 section). In the last option, a clustering strategy is applied to relate and group publications based on their semantic similarity. The terminology of different publications is compared, clustering the publications that shares similar terms. This way the method represent semantically similar publications grouped together.

3.4 Publication Annotation

Finally the proposed method allows the clinicians to annotate the results with two main functionalities: creating notes to remark any commentary about a

[4] `http://sphinxsearch.com/`

certain publication; and voting each publication relevance respect to the current query with like/dislike notations.

Actually all the notes are stored in the server database, in order to give the opportunity to clinicians to refer their already revised collections without repeating the entire process and showing the annotations they did in the past.

4 Method Validation

Two use cases have been developed within the environment of the *Tratamiento 2.0*[5] project to test the proposed method: diabetes and arterial hypertension. For each pathology, we have implemented a knowledge base with the relevant parameters. Based on these parameters, the service automatically extracts from EHR the characteristic of current patient, and translates it to the defined vocabulary. Concretely we have integrated English and Spanish version of the MeSH ontology at the NLP process for this validation. An example rule from the implemented rule-set is the one treating the systolic pressure (Sp) value: *if Sp <= 90: then "systolic hypotension"*.

A preliminary test set of ten patient data and fifty queries has been used to check the functionality of the method. The system has extracted correctly the characteristic for all the patients and it has correctly identified the 90% of the relevant papers, according to assessment of experts of the project.

In the context of the project *OSI+*[6], a user interface has been developed to exploit the modules defined above. With patient health record's characteristics and the current query, the searching criterion is defined. At this point the clinician has also to select which resources wants to use for searching literature and provide weights if a different importance among resources are identified. Nowadays, the *Cochrane library*, *PubMed* , *Fisterra* and *Ikere* are accessible resources from this implementation. The first two sources contain English literature while the last two mentioned resources store Spanish publications.

Based on the defined parametrization, the system searches for literature, and presents the results to the clinician. Previously launched searching results can be accessed, since the system maintains a cache of this information —not only the retrieved literature, but also the notes and the valuation the clinician has done for each publication. When previous searching results are consulted, this information is also presented to the clinician. The interface permits to search collection, not only using free text, but also exploiting the annotated terminology.

Currently we are working on integrating the clustered results visualization which is already running as service, but it is not exploited in the interface.

[5] `http://www.tratamiento20.es/inicio.html`

[6] `http://www.i3b.ibermatica.com/i3b/noticias/2009/osi-el-hospital-extendido`

5 Conclusions and Future Lines

In this paper we have presented a method to integrate biomedical literature repositories with patient EHRs. This method provides a platform for retrieving literature about a certain patient practice with minimal effort. The PICO query generation is supported, automatically extracting characteristics from patient EHR and applying NLP techniques for relevant terminology identification. Federated search strategy is used to access simultaneously to different information sources and to present a unified collection. With this method clinicians can edit the resultant collection, to record their considerations and to evaluate the relevance of each publication. All the information would be available afterwards to be accessed by the user without repeating the entire process. This method is also being standardized by using EHR formats such as HL7-CDA instead of a proprietary platform format.

Nowadays only access is provided to all the information generated by the users regarding the publications —i.e. agreement about the publications' relevance, tags and notes. But analysing that information, user profiles can be generated to facilitate model and adjustment of the results based on users' preferences. So, the method is intended to be the foundation of further development of a collaborative environments where clinicians could share their searches, notes and all the knowledge generated around a search. It would became a new resource of know-how that combines scientific studies with the daily experience of clinicians themselves.

References

1. Jimenez-Castellanos, A., Fernandez, I., Perez, D., Viejo, E., Díez, F.J., García de Kortazar, X., García, M., Maojo, V., Cobo, A: Patient based literature retrieval and integration: A use case for diabetes and arterial hypertension. Proceedings of the International Conference on Health Informatics, 33–41 (2011)
2. Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.E., Extracting information from textual documents in the electronic health record: a review of recent research. IMIA Year-book of Medical Informatics, Methods Inf Med 2008, 47 Suppl 1, 138–154 (2008)
3. Eichelberg, M., Aden, T., Riesmeier, J., Dogac, A., Laleci, G. B.: Electronic health record standards A brief overview. Information & Communications Technology (ICICT06) ITI4 (2006)
4. Dolin, R.H., Alschuler, L., Boyer, S., Beebe, C., Behlen, F.M., Biron, P.V., Shabo Shvo, A.: HL7 Clinical Document Architecture, Release 2. J. Am. Med. Inform. Assoc., vol. 13(1), 30–39 (2006)
5. Kalra, D., Beale, T., Heard, S.: The openEHR Foundation. Regional Health Economies and ICT Services: The PICNIC Experience, 115/2005, 153-173 (2005)
6. Lee, K. F.: Automatic speech recognition: The development of the SPHINX system. Kluwer Academic Pub (1989)
7. Liu, K., Mitchell, K.J., Chapman, W.W., Crowley, R.S.: Automating tissue bank annotation from pathology reports - comparison to a gold standard expert annotation set. Proceedings of the AMIA Annu Symp, 460-464 (2005)

8. Tsuruoka, Y., Tsujii, J., Ananiadou S.: FACTA: a text search engine for finding associated biomedical concepts. Bioinformatics vol. 24(21), 2559-2560 (2008)
9. Kim, J.J., Pezik, P., Rebholz-Schuhmann, D.: Medevi: Retrieving textual evidence of relations between biomedical concepts from medline. Bioinformatics, 24(11), 1410-1412 (2008)
10. Jacso, P.: Thoughts about federated searching. Information Today, 21(9), 17–20 (2004)
11. Schardt, C., Adams, M. B., Owens, T., Keitz, S., Fontelo, P.: Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med Inform Decis Mak, 7–16 (2007)
12. Meats, E., Brassey, J., Heneghan, C., Glasziou, P.: Using the Turning Research Into Practice (TRIP) database: how do clinicians really search?. J. Med Libr Assoc 156-163 (2007)
13. Huang, X., Lin, J., Demner-Fushman, D.: Evaluation of PICO as a knowledge representation for clinical questions. AMIA Annual Symposium proceedings, 359–363 (2006)