DISEArch
A Strategy for Searching Electronic Medical Health Records

David Elias Peña Clavijo, Alexandra Pomares Quimbaya and Rafael A. Gonzalez

Departamento de Ingeniería de Sistemas, Pontificia Universidad Javeriana, Bogotá, Colombia

Keywords: Medical Health Records, Health Data Mining, Text Mining.

Abstract: This paper proposes DISEArch, a novel strategy for searching electronic health records (EHR) of patients that have a specific disease. The objective of DISEArch is to enhance research activities on disease analysis allowing researchers to describe the disease they are interested on, and providing them the EHRs that best match their description. Its principle is to improve the precision of searching EHRs combining the analysis of structured attributes with the analysis of narrative text attributes producing a semantic ranking of EHRs with respect to a given disease. DISEArch is useful in medical systems where the information about the primary diagnosis of patients may be hidden in narrative text hindering the automatic detection of relevant records for clinical studies.

1 INTRODUCTION

Electronic health records (EHR) are a rich source of knowledge for medical research. However, their use has been limited due to the fact that important information is stored in narrative texts, intended for humans, difficult to search and analyse automatically. One of the requirements of medical research is to find the EHRs of patients that have been diagnosed with a specific disease. This task that should be easily done using classical queries (e.g. SQL) is very time-consuming because diagnosis is frequently hidden in the text (e.g. medical notes), hindering the possibility of automatically detecting relevant records and requiring the participation of an expert. Previous work on EHR systems propose strategies to improve automatic processing of narrative text in EHRs using information retrieval and data mining techniques (Han et al., 2006)(Zhou et al., 2005).

This work proposes DISEArch, a strategy for searching in EHRs those records that match a specific diagnosis, regardless of the kind of attribute (structured or non structured) that contains the information. DISEArch is composed of three phases. The first extracts the set of patient records from the medical health system. The second phase applies classical queries on structured attributes and text mining techniques over narrative text. Finally, it ranks the records by applying a semantic distance function with respect to the given disease description. DISEArch has been useful in reducing the time required for searching medical records. The structure of the paper is as follows. Section 2 presents the analysis of related works on narrative text and medical record analysis. Section 3 presents DISEArch, including its main components. Section 4 presents the main aspects of the prototype of DISEArch and the evaluation of its behaviour. Finally, Section 5 concludes this paper.

2 RELATED WORKS

Figure 1 presents a taxonomy of existing works related to text mining from EHRs. The initial categories offered are general approaches, algorithms, tools and scope. **General approaches** refer to three main bodies of work: information retrieval, natural language processing (NLP) and text/data mining. **Algorithms** are further divided into those aimed at data preparation and those aimed at data detection or classification. **Tools** offers a list of some available software tools which may support the process of text mining from EHRs. Finally, **scope** centers on work aimed at analyzing negated sentences, as opposed to work which is more generic. In our taxonomy (Figure 1) general approaches start with **text mining**, which consists of analyzing (portions of) documents typically made up of natural language. Its purpose is to uncover patterns, trends and relationships between words, meanings, terms or concepts (Spasic et al.,

Elias Peña Clavijo D., Pomares Quimbaya A. and A. Gonzalez R..
DISEArch - A Strategy for Searching Electronic Medical Health Records.
DOI: 10.5220/0004004601510156
In *Proceedings of the 14th International Conference on Enterprise Information Systems* (ICEIS-2012), pages 151-156
ISBN: 978-989-8565-10-5

Table 1: Comparative literature review.

	Algorithms								Tools				Approach		Field		
	Preparation			Classification/Detection													
Papers	Data transform.	Expert Tagging	UMLS Tagging	Regexp	Proposed	Bayesian networks	Decision trees	Hidden Markov M.	GATE	Link G. Parser	EMERSE	Own	Negated sentences	Generic	NLP	Data/text mining	IR
NegEx (Chapman et al., 2001)	√		√	√								√	√		√		
Context-Sensitive (Averbuch et al., 2004)	√		√		√	√	√					√	√				√
Negation-Recognition (Rokach et al., 2008)	√		√	√	√		√					√	√	√		√	
Geneneric Extraction (Han et al., 2006)			√						√	√				√	√	√	
DM & CBR (Huang et al., 2007)	√	√			√		√					√		√		√	
Text mining in biomedicine (Spasic et al., 2005)	√		√											√		√	√
Diabetic DW (Breault et al., 2002)	√	√					√					√		√		√	
EMERSE (Seyfried et al., 2009)		√									√			√			√
ABN (Antal et al., 2001)						√						√		√			√
Decision-Making (Claster et al., 2008)	√	√			√		√					√		√		√	
Semi-structured data to knowledge (Zhou et al., 2005)	√		√				√		√	√				√	√	√	√
HMM & LSA (Ginter et al., 2009)	√				√			√				√		√		√	√

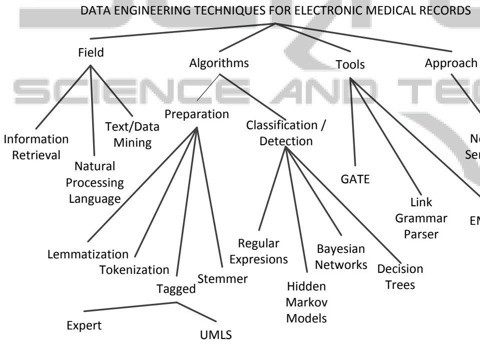

Figure 1: Taxonomy of EHRs data techniques.

2005). The second general approach is related to the field of **information retrieval (IR)**(Manning et al., 2008). The last general approach deemed useful for our purposes is **natural language processing (NLP)**, which refers to the recognition and use of information expressed in human language through computer-based systems (Hotho et al., 2005).

Our taxonomy continues by classifying specific types of algorithms that can be used as part of the three general approaches, depending on the stage of the process. With regards to **data preparation** we include four kinds of algorithms that prove useful in preparing unstructured health records prior to analysis.**Tokenization** is the process through which a flow of text is divided into segments. **Lemmatization** refers to a method in which verbs are transformed into their base form or nouns into their singular form. **Stemming** is used for removing irrelevant terms from the text. The last type of algorithm is **tagging**, which involves the interaction with a user that labels the text. In the case of EHRs, these tags are typically part of a controlled vocabulary, such as UMLS (USNLM, 2011). The second general type of algorithm in our taxonomy is grouped under **classification and detection**. **Decision trees** are a common part of the tool-belt for data mining and are useful in classifying conditions hierarchically such that a final decision is reached when a path can be followed from the root to one of its leaves. **Bayesian networks** are a powerful tool which is implemented through acyclic directed graphs that contain a set of nodes, each representing a random variable. (Antal et al., 2001).

A related kind of algorithm is called **hidden Markov model**, which represents a statistical model for linear problems and is widely used for speech recognition (Ginter et al., 2009). The third branch of related works is focused on the **tools** that support data and text mining in EHRs. Among these we find **GATE** (*General Architecture for text engineering*) which offers a general open source framework for developing or deploying software components for text engineering (Cunningham et al., 2011). Another tool is the **Link Grammar Parser**, which syntactically analyses text based on link grammar. **EMERSE** (*The Electronic Medical Record Search Engine*) (Hanauer, 2006) is specifically aimed at EHRs, acting as a search engine for free text inside such records. Using the taxonomy proposed above, Table 1 presents a comparative review of relevant literature around data / text mining in EHRs.

3 DISEArch STRATEGY

The strategies to examine narrative texts described in Section 2, provide a broad knowledge base to address the analysis of unstructured text inside EHRs. However, they are focused exclusively on the analysis of narrative text without taking into account the dependencies on other (structured) fields within the record. This work explores the combination of structured and narrative analysis to enhance the precision on the selection of relevant records for medical research. The strategy proposed in this section, called DISEArch, allows researchers to describe the disease they are interested in and provides them the set of health records that better match their description.

3.1 Phases

The process of analyzing health records in DISEArch is divided into the phases illustrated in Figure 2. In the first phase DISEArch allows medical researchers to describe the disease they are interested in using a template. This template includes formal and informal aspects of the disease, including the scientific name, the informal name, the tests that are typically used to diagnose the disease, and the symptoms of the disease. Once the disease is described, each of the fields are enriched using knowledge about the disease. The enrichment is made using a knowledge base created in OWL (Bechhofer et al., 2009) using a MeSH based thesaurus[1].

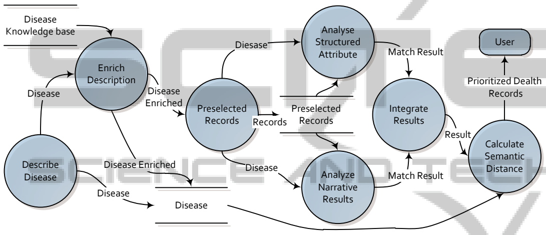

Figure 2: DISEArch Strategy.

The set of preliminary records are stored in a local database where DISEArch executes the analysis. The goal is to find the elements described in the disease within each one of the selected records, regardless of whether they are contained in an structured or a narrative text field. After this analysis, a score is given to each one and then prioritized.

3.2 Search Process

The disease description is divided into n subgroups S that are composed by m literals L (Definition 1). An example of subgroup is *Disease Name* and its literals are *Scientific name*, *Formal name*, *Informal name*, *Synonyms* and *Acronyms*. Similarly, health records M are divided into a set of p structured attributes S and q narrative text attributes T (Definition 2).

Definition 1 . ***Disease Description.*** *A Disease definition D is composed of a set of subgroups $S = \{s_1, s_2, ..., s_n\}$ that describe the main characteristics of the disease. Each subgroup s_i is specialized in a view of the disease and is composed of a set of literals $L(s_i) = \{l_{i1}, l_{i2}, ..., l_{im}\}$ where l_{ij} represents a fixed value for an atomic characteristic of the disease.*

Definition 2 . ***Health Record.*** *A health record M is composed of a set of structured attributes $C = \{c_1, c_2, ..., c_p\}$ whose domain of values is discrete and a set of narrative text attributes $T = \{t_1, t_2, ..., t_q\}$ whose domain is a natural language text.*

The goal of the search process is to detect within C_k and T_k of a record M_k, the value of each one of the literals l_{ij}. If the value of the literal l_{ij} is found in at least one attribute of the record M_k the value of the search process is changed to one (1), otherwise it is left at zero (0).

DISEArch contains two search functions in charge of detecting the occurrence of literal values into health records; the first one detects the value of a literal in structured attributes C and the second one searches within narrative text attributes T. Searching structured attributes is straightforward using classic sql queries. On the contrary, searching within narrative texts includes a previous preparation of texts and analysis that is detailed in Algorithm 1.

Algorithm 1: Narrative text search function.

```
Require: Record narrative text attributes
Ensure: Record search result
i,j,result ⇐ 0
for all textAttribute in record do
    $p$ ⇐ prepareText(textAttribute)
    for all subgroup in diseaseTemplate do
        for all literal in subgroup do
            ortResult ⇐ searchValue($p$).
            if $ortResult = 1$ then
                semResult ⇐ searchContext($p$)
                if $semResult = 1$ then
                    result ⇐ 1
                end if
            end if
            $j \Leftarrow j + 1$
        end for
        $i \Leftarrow i + 1$
    end for
    $recordResult[i, j]$ ⇐ result
end for
return recordResult
```

[1] http://www.nlm.nih.gov/

At the end of the search process the output is the score for each literal as the matrix *Res* illustrates and the number of hits for each one of the literals. The columns of *Res* represent the literals of each subgroup and the rows the health records.

$$Res = \begin{array}{c|cccccc} & l_{11} & l_{12}| & \cdots & |l_{n1} & l_{n2} & l_{n3} \\ \hline M_1 & 1 & 1 & \cdots & 1 & 1 & 1 \\ M_2 & 0 & 0 & \cdots & 1 & 1 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ M_k & 0 & 1 & \cdots & 1 & 0 & 0 \\ M_r & 1 & 0 & \cdots & 0 & 0 & 0 \end{array}$$

3.3 Integration Process

The integration process is in charge of representing the results of the search process taking into account the semantics of subgroups. This integration includes the results provided by structured and non-structured search functions. For doing this new representation the process takes into account the Definitions 3, 4, 5.

Definition 3 . ***Subgroup Intensity.*** *The intensity I of a subgroup of literals s_i is the normalised percentage of matched literals within the health record M_j. If a literal has multiple possible values (e.g. multiple acronyms) each value is considered a literal (e.g. Acronym 1, Acronym 2, etc.).*

Definition 4 . ***Subgroup Utility.*** *The utility U of a subgroup of literals s_i is a percentage value of the importance it has in identifying the disease diagnosed in a health record assuming that all the values of the literals are positive.*

Definition 5 . ***Subgroup Level of Hits.*** *The number of hits H of a subgroup s_i is the normalised number of times that literal values were matched within M_j.*

In order to calculate the utility of each subgroup DISEArch uses a classical method of multi-criteria decision analysis where each subgroup is evaluated on multiple criteria by experts and the utility is “the average specified in terms of normalised weightings for each criterion, as well as normalised scores for all options relative to each of the criteria” (Keeney and Raiffa, 1976). The number of hits is used as an optional calibration value that takes into account the number of times that literal values are found in a health record. The intention is to assign a higher weight to records that have the same literal multiple times. This value is optional because for some subgroups it is important, but for others it is not. At the end of the integration process an integration matrix is generated (see Matrix *I*). The values of the literal in each subgroup are described in the following columns:

1. s_i is 1 if at least one of the literals of the subgroup was found in the health record M_k, otherwise its value is 0.
2. s_i^I is the intensity of the subgroup.
3. s_i^U is the utility of the subgroup to detect the disease.
4. s_i^H is the number of hits of the subgroup. This columns is optional.

$$I = \begin{array}{c|cccccccc} & s_1 & s_1^I & s_1^U & s_1^H| & \cdots & |s_n & s_n^I & s_n^U \\ \hline M_1 & 1 & 1 & 0.6 & 1 & \cdots & 1 & 1 & 0.4 \\ M_2 & 0 & 0 & 0.6 & 0 & \cdots & 1 & 0.66 & 0.4 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ M_k & 1 & 0.5 & 0.6 & 0.5 & \cdots & 1 & 0.33 & 0.4 \\ M_r & 1 & 0.5 & 0.6 & 0.7 & \cdots & 0 & 0 & 0.4 \end{array}$$

The distance function between the disease description *D* and each one of the analyzed health records *M* is calculated using a distance function (e.g. Euclidean, Manhattan). The disease description as well as each record are represented in a n-space (see Function 1 and 2, respectively), where n is the number of subgroups.

$$D = (p_{s_1}, p_{s_2}, ..., p_{s_n}). \tag{1}$$

$$M = (q_{s_1}, q_{s_2}, ..., q_{s_n}). \tag{2}$$

The value of each point *p* is equivalent to s_i^U and the value of each point *q* is calculated using the product of $s_i^U \times s_i^I \times s_i^H$. The record with the shortest distance is the first one in the prioritized list and so on.

4 IMPLEMENTATION AND VALIDATION

In order to evaluate DISEArch and validate its improvement on the selection of the most relevant health records given a disease, a prototype has been constructed and used to evaluate its precision and recall. This section presents the main results obtained during this evaluation.

4.1 Prototype

For evaluating the behaviour of DISEArch we developed the components presented in Figure 3. These components are written in Java. The template of the disease can be filled using the GUI or directly using an XML file. The Dictionary Manager handles the knowledge base that allows the enrichment of the description of the disease. The knowledge base is implemented in OWL (Bechhofer et al., 2009). The Extraction Manager is in charge of the extraction and initial preprocessing of medical records from the EHR system. This component is parametrized according to the characteristics of the system and extracts the records according to the definition of initial parameters, such as date of admission, gender or age of patients. Persistent Manager and the DataStore Manager store the required tables to perform the search process inside a database. These tables are used to create a single view with all the unstructured and structured data. The component Text Mining is the core of the analysis and implements Stemming using Porter Stemmer algorithm, simple string tokenisation, sentence splitting, POS tagging using Probabilistic Part-of-Speech Tagging Using Decision Trees (Schmid, 1994) for annotating text with part-of-speech and lemma information and finally gazeteer lookup using regular expressions. This component has a coordinator that calls each of the search engines. The Narrative Search Engine is in charge of the analysis of natural language and was developed using the GATE API (Cunningham et al., 2011). This API enables the inclusion of all the language processing functionality within DISEArch. In addition, we use Treetagger (Schmid, 1994), a Pearl implementation which provides tokenization and Part of the Speech tagger. The Structured Search Engine is in charge of searching the disease over the structured attributes. Finally the Integrator component integrates the results using the semantic rules and prioritizes the set of records.

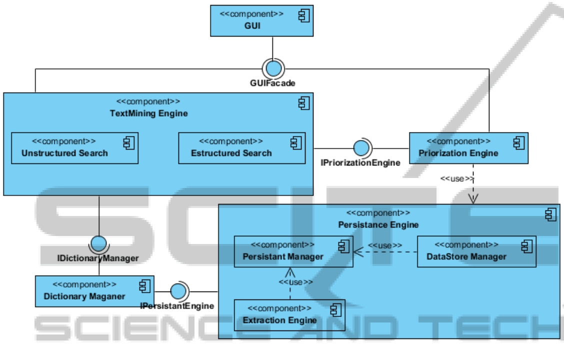

Figure 3: DISEArch component diagram.

4.2 Experiment Context and Results

Pulmonary Embolism (EP) was chosen to test DISEArch. A medical expert provided the subgroups and literals that describe it. Preliminary selection parameters for EHRs were defined: patients over 18 years old and records created between 2009-2011. One key item to obtain precision and recall was the prioritization process that was explained in Section 3. The results and their associated medical records were clustered according to their relevance (Lowly prioritized, Mildly prioritized and Highly prioritized medical records). The obtained results with DISEArch were 250 medical records, which correspond to records with at least one positive literal w.r.t the disease description. From these records, the prioritization process classified 30 as high, 52 as medium and 168 as low, according to the distance function.
In order to validate the precision and recall of DISEArch a medical expert analysed manually the records detecting 112 EP positive medical records. From these results DISEArch obtained 30 as high, 50 as medium and 32 as low. The precision and recall are presented in Figure 4. As expected the precision and recall is high for high and medium positive records. The low precision of the Low group is the consequence of the inclusion of records that contain few literals in common with the disease template.

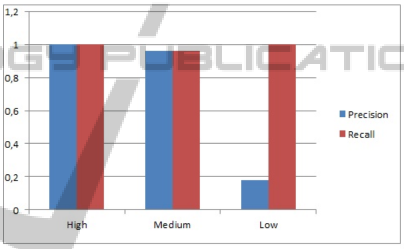

Figure 4: Precision and recall of DISEArch.

5 CONCLUSIONS

The DISEArch strategy presented in this paper enables medical researchers to identify those EHRs that include the diagnosis of a specific disease. This time-consuming and expert-dependent task can be supported by DISEArch through specific rules for identifying diseases and weights to prioritize the selected records, leaving the expert task to one of review and acceptance, rather than search and retrieval. DISEArch goes beyond classical text mining because it uses unstructured text in medical records as well as related structured fields to enrich the final results. From our first tests we found that, although the non-prioritized results are already helpful and accurate (as compared to expert selected records), prioritization still plays an important role in the classification of medical records because it adds precision and contributes to the review process by presenting the records in terms of how close they are to the disease template.

ACKNOWLEDGEMENTS

This work was supported by the project *“Identificación semiautomática de pacientes con enfermedades crónicas a partir de la exploración retrospectiva de las historias clínicas electrónicas registradas en el sistema SAHI del Hospital San Ignacio”* made by Pontificia Universidad Javeriana and Hospital Universitario San Ignacio.

REFERENCES

Antal, P., de Moor, B., and Mészáros, T. (2001). Annotated bayesian networks: A tool to integrate textual and probabilistic medical knowledge. In *Proc. of the 14th IEEE Symp. on Computer-Based Medical Systems*, CBMS ’01.

Averbuch, M., Karson, T. H., Ben-Ami, O., and Rokach, L. (2004). Context-sensitive medical information retrieval. *Studies in health technology and informatics*.

Bechhofer, S., van Harmelen, F., Hendler, J., and Horrocks, I. (2009). ”owl web ontology language reference”. Technical report, W3C.

Breault, J. L., Goodall, C. R., and Fos, P. J. (2002). Data mining a diabetic data warehouse. *Artificial Intelligence in Medicine*.

Chapman, W. W., Bridewell, W., Hanbury, P., and Cooper (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *J. of Biomedical Informatics*.

Claster, W., Shanmuganathan, S., and Ghotbi, N. (2008). Text mining of medical records for radiodiagnostic decision-making. *JCP*.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., and Aswani, N. (2011). *Text Processing with GATE (Version 6)*.

Ginter, F., Suominen, H., Pyysalo, S., and Salakoski, T. (2009). Combining hidden markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application. *I. J. Medical Informatics*.

Han, H., Choi, Y., Choi, Y. M., Zhou, X., and Brooks, A. D. (2006). A generic framework: From clinical notes to electronic medical records. *Computer-Based Medical Systems, IEEE Symp*.

Hanauer, D. A. (2006). Emerse: The electronic medical record search engine. *AMIA A. Symp Proc*.

Hotho, A., Nürnberger, A., and Paass, G. (2005). A brief survey of text mining. *LDV Forum*.

Huang, M.-J., Chen, M.-Y., and Lee (2007). Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. *Expert Syst. Appl.*

Keeney, R. and Raiffa, H. (1976). *Decisions with multiple objectives: Preferences and value tradeoffs*. J. Wiley, New York.

Manning, C. D., Raghavan, P., and Schtze, H. (2008). *Introduction to Information Retrieval*.

Rokach, L., Romano, R., and Maimon, O. (2008). Negation recognition in medical narrative reports. *I. R.*

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Seyfried, L., Hanauer, D. A., Nease, D., and Albeiruti (2009). Enhanced identification of eligibility for depression research using an electronic medical record search engine. *Inter. J. of Medical Informatics*.

Spasic, I., Ananiadou, S., McNaught, J., and Kumar, A. (2005). Text mining and ontologies in biomedicine: Making sense of raw text. *Briefings in Bioinformatics*.

USNLM (2011). Unified medical language system® (umls®). http://www.nlm.nih.gov/research/umls/. Noviembre 25, 2011.

Zhou, X., Han, H., Chankai, I., Prestrud, A. A., and Brooks, A. D. (2005). Converting semi-structured clinical medical records into information and knowledge. In *Proc. of the 21st Inter. C. on Data Eng. WS*.