2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)

Characterization of diseases based on phenotypic information through knowledge extraction using public sources

Gerardo Lagunes García
Centro de Tecnología Biomédica
Universidad Politécnica de Madrid
Pozuelo de Alarcón, Madrid, Spain
gerardo.lagunes@ctb.upm.es
ORCID: 0000-0002-6780-675X

Alejandro Rodríguez González
Centro de Tecnología Biomédica,
ETS Ingenieros Informáticos
Universidad Politécnica de Madrid
Madrid, Spain
alejandro.rg@upm.es
ORCID: 0000-0001-8801-4762

***Abstract*— Despite the huge findings made by the study of the behaviour of diseases, there are currently many non-cure or non-treatment diseases and only some of their symptoms can be beaten. Understanding how the diseases behave implies a complex analysis that together with the new technologies provide researchers with more calculation and observational capabilities, as well as novel approaches that allow us to observe how the diseases behave and relate in different environments with distinct factors. Current research aims to find new ways of characterizing the diseases based on phenotypic manifestations using knowledge extraction techniques from public sources. With the characterization of the diseases, a better understanding about the diseases and how similar they are can be achieved, leading for example to find new drugs that can be applied to different diseases. In order to carry out the present research we have made use of our own dataset of symptoms and diseases developed using an approach that allows us to generate phenotypic knowledge from the extraction of medical information from several data sources.**

Keywords—medical knowledge extraction, disease characterization, human symptoms-disease dataset

I. The Research Problem

The complexity of the study of the several diseases is well known, as there are many factors that are involved in each of them: biological, genetic, environmental, physical factors, among others. The Pathology is the field of the natural sciences that studies the diseases, being the study and its practical application the fundamental means to know the objective reality and to acquire knowledge.

The knowledge acquired implies to know, interpret, transform and apply this scientific knowledge, using as a fundamental premise that without research there is no future. There are currently between 6,000 and 7,000 rare diseases (in that they affect one person out of every 2,000), for which there is no treatment exactly because they are not known [1], [2].

In order to gain a better understanding of them, scientists try to categorize them, that is, to create large groups of diseases by the common elements between them, thus, in 2007 the Human Disease Network (HDN) was created [3], a network that categorises diseases by their genetic elements; later, in 2014, the HSDN was created, which tries to categorise diseases by their phenotypic similarities [4]. Through the HSDN, it could be said that, in many cases, those diseases that shared phenotypic similarities also shared genetic elements.

Our work aims to go beyond the study of diseases as such, while the previous networks (HDN and HSDN) have extracted information from OMIM and PubMed, respectively, in this case, our approach aim is to obtain as much information as possible from public information sources; Wikipedia, PubMed and MayoClinic having been mined, at the moment.

In spite of the so reduced value that they are usually given in sources like Wikipedia, for the fact of being within reach of anyone, even those people who have limited or null knowledge in medicine, the truth is that this source constitutes an important medical library.

The main advantage of the selected sources is the continuous updating to which they are exposed due, precisely, to the free and open access to them, which allows all people who can Internet access can improve the information contained in them, generating a comprehensible disease-and-symptom dataset. And on this collection of diseases, the verifications of the hypotheses described in the section "Hypotheses" will be carried out.

II. State of the Art

To improve their knowledge about diseases and to understand how they are related and how they behave, health professionals need quality, integrated, highly available and as structured data as possible. For example, the Goh et al study that related disorders with genes and phenotypes with genes reveals that *essential human genes are likely to encode hub proteins and are expressed widely in most tissues, suggesting that disease genes also would play a central role in the human interactome* and for that built an HDN with gene and disorder data from OMIM but the built dataset is not available [3]. In the same context to build the HSDN its developers extract information from MeSH and PubMed, and with its study discovered that *diseases that share symptoms indicate shared protein interactions* but again the built dataset is not available [4]. In the line of genes, DisGeNet aims to form a

DOI 10.1109/CBMS.2019.00124

comprehensive resource available on diseases and their genes from several knowledge sources, also provide a set of analysis tools to facilitate and foster the study of the molecular underpinning of human diseases [5]. Biomedical knowledge is also included in free access ontologies or vocabularies such as UMLS [6], HPO [7], DO [8], among others. Some projects that have created freely available disease databases are; OMIM a curated knowledge base of human genes, phenotypes and genetic disorders [9]; MalaCards which is called the human disease database that includes knowledge extracted from 68 distinct sources and the GeneCards, a sister project that focuses on human gene information [10], [11]; DiseaseCard is another endeavour that links genetic and medical information from other sources on rare genetic diseases to facilitate navigation through the different sources [12]; the DISEASES system produced a compendium of disease-gene associations by extracting them from Medline abstracts. On the other hand, there are resources such as the Diseases Database (DD), which despite valuable content is not programmatically accessible. Finally, the analysis of the related works shows interesting approaches to apply and improve for the construction of the disease cluster. First, the use of multiple sources of biomedical information, second, provide structured data and finally share knowledge through a REST API.

III. Hypothesis

The hypotheses that lead to the work presented in this thesis are:

- Hypothesis 1 (**H1**): It is possible to extract and generate accurate phenotypic information from public sources using NLP-based approaches (specifically MetaMap).
- Hypothesis 2 (**H2**): The phenotypic information contained in Wikipedia articles categorized as diseases is of quality.
- Hypothesis 3 (**H3**): As time goes on, these sources contain more or better information leading to an enrichment of knowledge about diseases.

IV. Material and Methods

As a means of verifying the hypotheses proposed, a dataset of diseases and symptoms has been created using public sources of medical information. This section describes in detail the means used and the pipeline that shapes the dataset previously mentioned.

A. Source Selection and Disease List

The acquisition of knowledge about diseases was traditionally found in books or manuals of medicine, valuable elements, but with limitations: they are not updated frequently, they are not within reach, the automatic access to them is complex and due to their learning purpose, their content is not structured for mining tasks. Currently, the Web contains free biomedical resources available to people with Internet access. Biomedical Web knowledge is found in three kinds of information sources, first in abstracts and in some cases in the full text of medical articles from platforms such as PubMed; second, in information sources like MayoClinic, MedlinePlus, CDC, Orphanet, among others and finally in articles from non-specialized sources like Wikipedia or Freebase.

The requirements when selecting a source of knowledge are: **a)** open access, **b)** recognised quality and reliability, and **c)** availability of substantial amounts of data (structured or unstructured). Therefore, this criterion allows incorporating three first sources: **i)** Wikipedia, **ii)** PubMed, and **iii)** MayoClinic. The multi-source approach implies the capacity to add new knowledge from new sources.

Another differentiator is the use of Wikipedia as a primary source of medical information retrieval. Despite being a valid medical encyclopaedia [13]–[17], it has not been used to obtain updated and rich mix knowledge by health professionals in one place about the signs and symptoms of diseases, which provides an opportunity to analyse the quality of their diagnostic knowledge [14], [18]. The approach to create the list of diseases is to obtain from DBpedia [19] and and DBpedia-Live[20] (with structured information from Wikipedia) through a SPARQL query all articles tagged as diseases and select only those containing sections identified with phenotypic content: *Signs and symptoms*, *signs and symptoms*, *Symptoms and causes*, *Signs*, *Symptoms*, *Causes*, *Cause*, *Diagnosis*, *Diagnostic*, *Causes of injury*, *Diagnostic approach*, *Presentation*, *Symptoms of …* , *Causes of …* , and *infobox*. In the *infobox* section you will find several disease identifiers in external information sources like MeSH, OMIM, MedlinePlus, ICD-10, DiseasesDB, among others; these identifiers, if detected, allow to select relevant articles.

MayoClinic (www.mayoclinic.org) is a nonprofit organization committed to education, research and clinical practice. In USA it is considered one of the best providers of health services[1, 2], and in the field of research has a long history of scientific publications, thereby has a quality medical database on diseases and their symptoms. On their website they have a list of diseases, which we integrate to our system. Each article associated with a disease has sections that contain a general description of the disease, how it is presented, its causes, diagnoses, treatments and the kinds of doctors who treat it in their departments. The sections of interest that focus on phenotypic content are: *Symptoms*, *Causes* and *Diagnostic*.

The last source of information selected is PubMed, a valuable medical source for its scientific importance in the field of medicine and houses millions of medical articles. Despite the large textual content of this platform, has the limitation that you cannot access the full text of a large number of their articles and sometimes not even have access to the abstract. To mine it is necessary first, to obtain the list of medical terms related to human diseases from the MeSH classification tree (from the C01 to the C20 section); second, through PubMed's Entrez API to obtain the 100 most relevant articles from each of the previously selected MeSH terms. For each article we recover: the abstract, authors' names, unique

[1] https://www.mayoclinic.org/es-es/about-mayo-clinic/quality/rankings

[2] https://www.mayoclinic.org/es-es/about-mayo-clinic/office-diversity-inclusion

identifier in PubMed and PubMed Central, doi (digital object identifier), title, associated MeSH terms and keywords [21].

All this diversity of sources brings with it the challenge to mine diverse characteristics in structure, writing, comprehension and actuality.

B. Data Retrieval and Knowledge Extraction Approach

The first step to develop the symptoms–diseases dataset is in charge of retrieving the information from the sources previously identified and described. For each one of this, and before running the actual web crawler, the *Get Disease List Procedure* (GDLP) component is responsible for obtaining the list of diseases to be mined, thus providing links to all available disease related documents. For example, the GLDP associated to Wikipedia articles makes use of the SPARQL query; similarly, the links for the PubMed's articles are retrieved through a list of MeSH terms. However, in the case of MayoClinic, the terms are retrieved by scrapping strategies.

Once the URL list has been collected, the *Web Crawler* (WC) module is in charge of connecting to each of the hyperlinks and extracting the specific text that describes the phenotypical manifestations, as well as the links (references) contained within the texts. In addition, and whenever possible, it attempts to extract information related to the coding of diseases, i.e. the codes used to identify the disease in different databases or existing data vocabularies. Currently it is able to retrieve information from more than 5,500 articles in Wikipedia, from 229,160 article abstracts in PubMed and from 1,180 articles in MayoClinic. The information mined by WC is stored in an intermediate database called *Raw DB*, which contains the raw unprocessed text.

The next step within the pipeline is called *NLP Process* (NLPP). This component is responsible for: **i)** reading all the texts of a snapshot, and **ii)** obtaining for each text a list of relevant clinical concepts/terms, discarding any unrelated paragraphs or words. At the moment NLPP uses Metamap [22][23] as a Natural Medical Language Processing tool to extract clinical terms of interest – see online NLP Tools and Configuration section[3].

The output of the NLP process is stored in the *Medical DB* (DMDB) database. It stores, in a structured way, the medical concepts that have been obtained by the NLPP, as well as any information required to track the origin of such concepts – in order to track any error that may later be detected. Therefore, and to summarize, the information stored in a structured way in DMDB is: **i)** the medical concepts with their location, information and semantic types, **ii)** the texts from which they were extracted and the links by them contained, **iii)** the sections which the texts belong to, **iv)** the document or documents describing the disease (Web link) and **v)** the disease identifiers codes in different vocabulary or databases. Additional information, as the day of the extraction and the source, is further saved.

Having clarified this, the next component of the pipeline, the *TVP Process* TVPP, reads all the concepts of a snapshot - source pair and filters them. This process is responsible for determining whether these UMLS medical terms are really phenotypic manifestations, and for storing the results back in the DMDB. TVPP is based on the Validation Terms Extraction Procedure that was developed, implemented and tested by Rodriguez-Gonzalez et al [24]. The results of this component (a purification of concepts) are thus those validated terms that we will consider as true phenotypic manifestations of diseases. For a better tracking of our data retrieval and knowledge extraction approach, online is the diagram with the workflow[4].

The *Extraction Process* (IEPD), i.e. the process of retrieving and storing information about diseases, basically ends here. Nevertheless, for the sake of providing an accessible and user-friendly way of retrieving and manipulating this information, are also offers a REST-based interface, whose documentation is online (http://disnet.ctb.upm.es/apis/disnet).

V. Validation

This section describes the validation methods for each of the presented hypotheses.

The **H1** has been developed to verify the extraction and knowledge generation method from public sources; and for this we will perform a manual validation of the phenotypic knowledge generated by our approach. The **H2** aims to verify the quality and confidence of the texts with phenotypic content of Wikipedia articles tagged as human diseases; therefore, we will compare the Wikipedia disease-symptom dataset in contrast to other reliable sources (e.g.: MedlinePlus, Diseases Database, Malacards). And finally, the **H3** allows to verify if the evolution of the data in the public sources implies an enrichment in the knowledge we have about the diseases; and to know it we will verify if the data change and we will validate if this increase represents a real knowledge.

VI. Preliminary Results

A repository of human symptoms-diseases was built with data extracted from three sources of information: Wikipedia, PubMed y MayoClinic. From Wikipedia we have 24 snapshots, from February 1st, 2018 to February 1st, 2019, for PubMed we have one snapshot, that of April 3, 2018 and for MayoClinic we have 12 snapshots, from August 15th, 2018 to February 1st, 2019. All snapshots were created using the same Metamap configuration[3]. Within the system it is possible to consult, for each snapshot and source, the total number of articles with medical terms, the total number of medical terms found, the number of processed texts, the total number of retrieved codes, and the total number of semantic types found[5]. At the latest snapshot Wikipedia has 4,775 diseases, PubMed 2,212 diseases and MayoClinic 1,124 diseases. The dataset also has

[3] http://disnet.ctb.upm.es/apis/disnet#NLP_Tools_and_Configuration

[4] https://midas.ctb.upm.es/gitlab/disnet/paperdisnet/blob/master/final_disnet_work_flow_v0.2.png

[5] https://midas.ctb.upm.es/gitlab/disnet/paperdisnet/tree/master/knowledge_sources

information on more than 2,211 UMLS medical terms tentatively as symptoms and signs, divided into 16 semantic types.

The evaluation approach has been tested by executing the prototype over data from 100 different Wikipedia diseases and 100 different PubMed abstracts selected randomly with the only filter that diseases have at least 20 medical terms. The evaluation was performed by doing a manual analysis of the results provided by our approach. For each disease, we compared: **(1)** the list of terms provided by our approach; with: **(2)** a list of terms manually extracted from the disease Web page. True positive (**TP**), false positive (**FP**), true negative (**TN**) and false negative (**FN**) parameters were computed in order to calculate precision, recall, specificity and F1 score values. Results indicate that our NLP (Metamap + TVP) process is sufficiently reliable, with an accuracy of 0.753 (confidence interval of (0.730, 0.775)) for Wikipedia and of 0.644 (confidence interval of: (0.606, 0.680)) for PubMed.

ACKNOWLEDGMENT

This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No. 727658, project IASIS (Integration and analysis of heterogeneous big data for precision medicine and suggested treatments for different types of patients). The paper is also a result of the project "DISNET (Creation and analysis of disease networks for drug repurposing from heterogeneous data sources applied to rare diseases)", that is being developed under grant "RTI2018-094576-A-I00" from the Spanish Ministerio de Ciencia, Innovación y Universidades. Gerardo Lagunes-Garcia work is supported by Mexican Consejo Nacional de Ciencia y Tecnología (CONACYT) (CVU: 340523) under the programme "291114 - BECAS CONACYT AL EXTRANJERO".

REFERENCES

[1] F. U. and D. DeGette, "Can we find cures for 7,000 diseases? (Opinion)," CNN. [Online]. Available: https://www.cnn.com/2015/01/13/opinion/upton-degette-cure-diseases/index.html. [Accessed: 12-Jun-2018].

[2] https://www.facebook.com/FactChecker, "Are there really 10,000 diseases and just 500 'cures'?," Washington Post. [Online]. Available: https://www.washingtonpost.com/news/fact-checker/wp/2016/11/17/are-there-really-10000-diseases-and-500-cures/. [Accessed: 26-Jan-2019].

[3] F. Emmert-Streib, S. Tripathi, R. de M. Simoes, A. F. Hawwa, and M. Dehmer, "The human disease network," Syst. Biomed., vol. 1, no. 1, pp. 20–28, Jan. 2013.

[4] X. Zhou, J. Menche, A.-L. Barabási, and A. Sharma, "Human symptoms-disease network," Nat. Commun., vol. 5, p. 4212, Jun. 2014.

[5] J. Piñero et al., "DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes," Database, vol. 2015, Jan. 2015.

[6] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," Nucleic Acids Res., vol. 32, no. suppl_1, pp. D267–D270, Jan. 2004.

[7] P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos, "The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease," Am. J. Hum. Genet., vol. 83, no. 5, pp. 610–615, Nov. 2008.

[8] L. M. Schriml et al., "Disease Ontology: a backbone for disease semantic integration," Nucleic Acids Res., vol. 40, no. Database issue, pp. D940–D946, Jan. 2012.

[9] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," Nucleic Acids Res., vol. 33, no. Database Issue, pp. D514–D517, Jan. 2005.

[10] N. Rappaport et al., "MalaCards: A Comprehensive Automatically-Mined Database of Human Diseases," Curr. Protoc. Bioinforma., vol. 47, no. 1, pp. 1.24.1-1.24.19, Sep. 2014.

[11] S. Espe, "Malacards: The Human Disease Database," J. Med. Libr. Assoc. JMLA, vol. 106, no. 1, pp. 140–141, Jan. 2018.

[12] J. L. Oliveira et al., "DiseaseCard: A Web-Based Tool for the Collaborative Integration of Genetic and Medical Information," in Biological and Medical Data Analysis, 2004, pp. 409–417.

[13] T. Shafee, G. Masukume, L. Kipersztok, D. Das, M. Häggström, and J. Heilman, "Evolution of Wikipedia's medical content: past, present and future," J Epidemiol Community Health, p. jech-2016-208601, Aug. 2017.

[14] A. Azzam et al., "Why Medical Schools Should Embrace Wikipedia: Final-Year Medical Student Contributions to Wikipedia Articles for Academic Credit at One School," Acad. Med., vol. 92, no. 2, pp. 194–200, Feb. 2017.

[15] J. M. Heilman and A. G. West, "Wikipedia and Medicine: Quantifying Readership, Editors, and the Significance of Natural Language," J. Med. Internet Res., vol. 17, no. 3, Mar. 2015.

[16] N. Thompson and D. Hanley, "Science Is Shaped by Wikipedia: Evidence from a Randomized Control Trial," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3039505, Sep. 2017.

[17] D. Matheson and C. Matheson, "Open Medicine Journal Wikipedia as Informal Self-Education for Clinical Decision-Making in Medical Practice," Open Med. J., vol. 4, pp. 1–25, Sep. 2017.

[18] N. Cohen, "Editing Wikipedia Pages for Med School Credit," The New York Times, 29-Sep-2013.

[19] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A Nucleus for a Web of Open Data," in The Semantic Web, Springer, Berlin, Heidelberg, 2007, pp. 722–735.

[20] M. Morsey, J. Lehmann, S. Auer, C. Stadler, and S. Hellmann, "DBpedia and the live extraction of structured data from Wikipedia," Program, vol. 46, no. 2, pp. 157–181, Apr. 2012.

[21] D. Westergaard, H.-H. Stærfeldt, C. Tønsberg, L. J. Jensen, and S. Brunak, "A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts," PLOS Comput. Biol., vol. 14, no. 2, p. e1005962, Feb. 2018.

[22] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program," Proc. AMIA Symp., pp. 17–21, 2001.

[23] A. Rodríguez González, R. Costumero Moreno, M. Martínez Romero, M. D. Wilkinson, and E. Menasalvas Ruiz, "Extracting diagnostic knowledge from MedLine Plus: a comparison between MetaMap and cTAKES Approaches," Curr. Bioinforma., vol. 375, no. 0, pp. 1–7, 2015.

[24] A. Rodríguez-González, M. Martínez-Romero, R. Costumero, M. D. Wilkinson, and E. Menasalvas-Ruiz, "Diagnostic Knowledge Extraction from MedlinePlus: An Application for Infectious Diseases," in 9th International Conference on Practical Applications of Computational Biology and Bioinformatics, Springer, Cham, 2015, pp. 79–87.