Assisting Data Retrieval with a Drug Knowledge Graph

Romain LELONG[a,b,1], Badisse DAHAMNA[a,b], Romain LEGUILLON[a], Julien GROSJEAN[a,b], Catherine LETORD[a,b], Stéfan J.DARMONI[a,b], Lina F. SOUALMIA[b,c]
[a]*CHU Rouen, Department of Biomedical Informatics, F-76000 Rouen, France*
[b]*LIMICS U1142, Sorbonne Université, Paris, France*
[c]*Normandie Univ, UNIROUEN, TIBS-LITIS EA 4108, F-76000 Rouen, France*

Abstract. The Normandy health data warehouse EDSaN integrates the medication orders from the University Hospital of Rouen (France). This study aims at describing the design and the evaluation of an information retrieval system founded on a complex and semantically augmented knowledge graph dedicated to EDSaN drugs' prescriptions. The system is intended to help the selection of drugs in the search process by health professionals. The manual evaluation of the relevance of the returned drugs showed encouraging results as expected. A deeper analysis in order to improve the ranking method is needed and will be performed in a future work.

Keywords. Drug Information Retrieval; Knowledge Graph; Semantic Network.

1. Introduction

Data provided and/or exploited by drugs systems usually fall into two types: (a) factual drugs data and (b) knowledge drugs data. Factual drugs data mainly consist in drugs prescription and drugs administration data that are archived by hospitals usually as free text within the discharge letter or medication orders. Several methods have been proposed to perform information retrieval on factual drugs data: information extraction and free text search [1], machine learning [2]. However, the implementation of an effective information retrieval system requires the use of knowledge data in addition to factual data. Knowledge graph structure including the conceptual graph formalism [3] has been used for biomedical knowledge and data representation is particularly suited to drug knowledge data [4]. Existing medicinal drug databases such as Wikidata [5], Drug Bank[2], or GoodRx[3] contain valuable information but lack of comprehensiveness when taken separately and/or store some of this information as unstructured data [6]. In this study, the design of a system enabling the retrieval of prescription orders contained in the Normandy's Health Data Warehouse (EDSaN) [7] at the Rouen University Hospital (Normandy, France) is described. A conceptual graph of drug knowledge data was designed and used in the information retrieval process to retrieve the French Common

[1] Corr. Author, Department of Biomedical Informatics, Rouen University Hospital, 1 rue de Germont Rouen France; E-mail: romain.lelong@chu-rouen.fr.

[2] https://go.drugbank.com/releases/latest

[3] https://www.goodrx.com/

Dispensing Unit (UCD) codes that are used to encode and bill administrated drugs in France. A first evaluation is conducted to assess the consistency of the concepts resulting from traversing the graph. The evaluation focused in this study on the UCD codes only.

2. Material and Method

2.1. The EDSaN Data Warehouse and HeTOP

The EDSaN data warehouse currently integrates 6,978,586 atomic prescriptions, *i.e.* of a single pharmaceutical specialty, distributed over 1,452,616 prescriptions' orders between 2011 and the end of August 2021. The data are loaded and maintained into EDSaN from the Rouen University Hospital Information System (HIS) and originate from either a dedicated database or in the care plan. Although various structured fields exist in the HIS to describe the prescriptions and administrations of drugs (dosage, time, number of administrations, etc.), the poor quality of these data, e.g. empty fields, wrong values, does not allow an accurate structured search. In EDSaN, every atomic prescription is associated to a UCD code that identifies the prescribed pharmaceutical product. The prescription orders include patient information (id, age, gender, and birthdate), stay information (id, entry and leaving dates, units) and prescription information (id, date), and a list of atomic prescriptions that corresponds to a single prescribed drug. The Health Terminology/Ontology Portal (HeTOP) (https://hetop.eu) is used as primary data source for drug knowledge data. It currently integrates terminological concepts from over than 70 health terminologies and/or ontologies as well as semantic relationships between those concepts. Since 2019, it includes a formal drug model suited for French specificities.

2.2. Structured Information Retrieval

The free and open-source search engine software library Apache Lucene was used to enable the querying of prescription orders. Lucene documents were generated for each order to provide Boolean search functionalities. A graphical module specifically dedicated to prescriptions were also added to the Java web application of EDSaN. It enables users to query the Lucene indexes through a form that provides a search field for each prescription order metadata. This form enables an accurate structured search since each metadata could be targeted. The search for prescription orders based on what drug were prescribed can notably be done by using the search field dedicated to UCD codes. This is essential to handle user queries for which the full text search does not return any answers. However, this task requires a pre-selection of the UCD codes to be targeted.

2.3. Modelling the Knowledge Graph

To assist in the selection of UCD codes, a knowledge graph was designed thanks to drug and related data (e.g., diseases) from HeTOP and its semantic network. Vertices are selected among HeTOP concepts and include MeSH (Medical Subject Heading) Descriptors, Anatomical Therapeutic Chemical (ATC) concepts, International Nonproprietary Names (INNs), virtual drugs from the MédicaBase (https://www.medicabase.fr/index.htm) database, pharmacological roots, pharmacological specialties, drug components, drug components groups, and UCD codes. The edges between vertices were also taken from semantic relationships in the HeTOP. Some types of relationships were excluded due to their technical nature or low relevance in the context of this study (e.g., “not to be confused with”). The overall drug knowledge graph is composed of 131,277 vertices and 703,807 edges. When possible, the user's input is matched to a vertex of this graph. The paths starting from that node and leading to UCD codes are then aggregated and proposed to the user in the form of a selectable sub-graph. In order to provide the user with the most relevant UCD codes first, the paths resulting from the user's query are ranked according to a weight and then filtered. The weight of a path is calculated by summing the weights of the relationships traversed in that path. The weights of the relationships were assigned empirically in collaboration with a hospital pharmacist. Seven ascending ranges of path weight were defined in order to distribute the paths over seven corresponding relevance classes (C1 to C7). In the interface, a slider component allows the user to adjust the tolerance and then show or hide the paths assigned to these different classes in the resulted graph (see Figure 1).

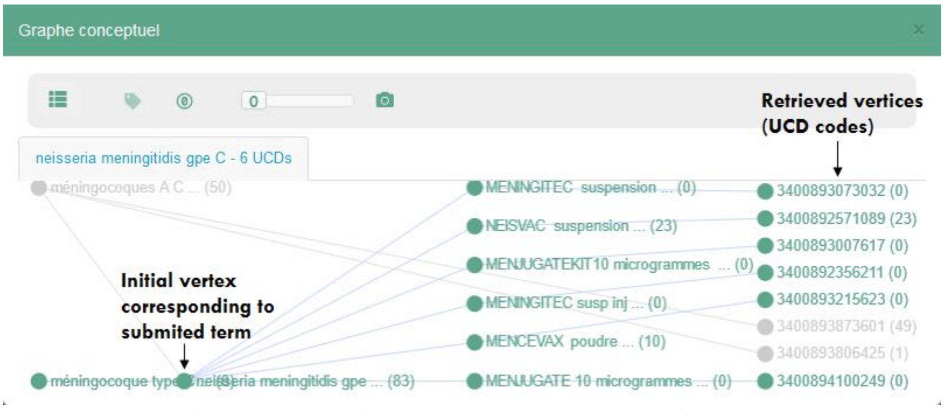
Graphe conceptuel
0
Retrieved vertices (UCD codes)
neisseria meningitidis gpe C - 6 UCDs
méningocoques A C ... (50)
MENINGITEC suspension ... (0)
NEISVAC suspension ... (23)
MENJUGATEKIT 10 microgrammes ... (0)
MENINGITEC susp inj ... (0)
MENCEVAX poudre ... (10)
MENJUGATE 10 microgrammes ... (0)
3400893073032 (0)
3400892571089 (23)
3400893007617 (0)
3400892356211 (0)
3400893215623 (0)
3400893873601 (49)
3400893806425 (1)
3400894100249 (0)
Initial vertex corresponding to submited term
méningocoque type C (8)
neisseria meningitidis gpe ... (83)

Figure 1. Example of a path traversal of the knowledge graph at a Class 1 from the user query “neisseria meningitidis gpe C” which corresponds to INN 11565. It allows the proposal of paths leading to height relevant UCDs. Three types of paths were browsed in this example: INN > pharmacological specialty > UCD Code or INN > ATC Code > UCD Code or INN > ATC Code > pharmacological specialty > UCD Code.

3. Results

To evaluate the semantic graph modelling the drug knowledge, a set of n=88 terms was randomly drawn among the possible types of vertex. These terms led to more than 100,000 paths overall. For each term, only the UCD codes issuing from paths belonging to the first three non-empty relevance class proposed by the system were evaluated. This UCD codes assessment was done by a single expert, a hospital pharmacist and consisted in giving a score: 1 if the UCD code seemed to him unsatisfactory, 2 if it could be improved, and 3 if the result was consistent. At this evaluation stage, neither the path itself, nor the relevance class assigned to it by the system were considered. Table 1 summarizes the results of the evaluation.

Table 1. Average of scores given by the expert for the first non-empty relevance classes from C1 (min path weight) to C3 (deeper path). Nt: number of terms; Np: number of paths obtained. The shaded classes have not been taken into account in this study.

Type of term	Nt	Np	C1	C2	C3	C4	C5
drug components	7	1,514	**2.71**	1.08	1.04	▨	▨
drug components groups	3	2,270	-	**3.00**	1.00	1.00	▨
MeSH Descriptors	11	20,665	-	**2.94**	1.49	1.05	▨
INN	7	2,232	-	-	**1.95**	1.01	1.00
virtual drugs	11	2,470	**3.00**	1.23	1.28	▨	▨
Medical indication	6	3,341	-	-	**3.00**	2.82	1.52
ATC Code	14	79,094	2.25	-	**3.00**	1.00	▨
pharmacological specialties	7	3,198	**2.50**	1.59	1.89	▨	▨
pharmacological roots	22	4,648	**3.00**	1.84	1.05	▨	▨

4. Discussion and Conclusion

As the assignments of relevance classes is based on relationships weights that have been empirically chosen, the average score for each class needs to be assessed. From a general point of view, one can observe that for each type of term, the average scores of the paths tends to decrease with the level of relevance (i.e. the class) assigned to it. Although, exceptions can be found, especially with regards to the ATC terms (C1=2.25 whereas C3=3.00). This shows that the weights of the relationships have been consistently assigned. Nevertheless, significant variability and inconsistency can be observed between term types. The highest score of 3 is reached for several types of terms (e.g. Drug Component Groups (C2=3), Virtual Drugs (C1=3). One can observe that for Medical Indications and MeSH Descriptors, their score is still high, despite the fact that these terms are generic and consequently harder to semantically link to UCD codes than pharmacological specialties or ATC codes for instance. Moreover, the best possible relevance class that can be obtained varies among the types of terms (C3 for Medical Indications and INNs, C2 for drug components groups and MeSH descriptors, C1 for others) although the score of those class remain close to 3.00.

The main aim of the drug knowledge graph was to assist the user in selecting drugs of interest. From that perspective, this study showed encouraging results as the ranking of resulting UCDs codes were overall congruent with the expert judgment. However, some inconsistencies remain. Future work will therefore focus on the refinement of the weights assigned to the edges of the graph.

References

[1] Dietrich G *et al.* Replicating medication trend studies using ad hoc information extraction in a clinical data warehouse. *BMC Med Inform Decis Mak. 2019 Jan 18;19(1):15*.

[2] Corny J *et al.* A machine learning–based clinical decision support system to identify prescriptions with a high risk of medication error. *J Am Med Inform Assoc. 2020 Nov 1;27(11):1688-1694*.

[3] Sowa JF. *Conceptual structures: information processing in mind and machine*. Addison-Wesley Longman Publishing Co., Inc., 1984.

[4] Teng A, et al. Biomedical Graph Visualizer for Identifying Drug Candidates. *bioRxiv*, 2020.

[5] Vrandečić D & Krötzsch M. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78-85.

[6] Mann M, *et al*. Open Drug Knowledge Graph. 2021.

[7] Lelong R, *et al*. Building a semantic health data warehouse in the context of clinical trials: development and usability study. JMIR Med Inform. 2019;7(4):e13917.