October 2018

Measuring distances between medical entities. Step 1: DrugBank

Alberto OLIVARES-ALARCOS [a,1], Iva STANKOVIC [a], Humberto GONZÁLEZ [a] and Horacio RODRÍGUEZ [a]

[a] *Department of Computing Science, Universitat Politècnica de Catalunya, UPC*

Abstract. We face in this paper the problem of computing distance measures between medical entities. Specifically we deal with the most challenging type of medical entity: drugs. Three different similarity measures between drugs are presented, based each one on specific dimensions of drugs description, namely textual, taxonomic and molecular information. All the information has been extracted from the same resource, the *DrugBank* database.

Keywords. Distance Measures, Medical entities, Drugs, DrugBank

1. Introduction

More than one century ago, Lord Kelvin (William Thomson) said "I often say that when you can measure what you are speaking about and express it in numbers you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind". This quote summarizes our objective in the research reported here. We are working on the Biomedical domain and we are interested in putting numbers on the medical entities. The most operative way of doing so is to compute the distance[2] between medical entities. There are many types of medical entities, diseases, body parts, drugs, symptoms, medical findings, etc.[3]. We have chosen drugs as our first step on our aim because its interest, its difficulty, and the availability of a comprehensive lexical resource, *DrugBank*[4] [1], as a semantic space for applying the metrics.

The ways of computing distances between drugs basically are based on some intuitions:

- Two drugs are similar if their names are similar.
- Two drugs are similar if their descriptions are similar.
- Two drugs are similar if they are frequently applied over similar targets.

[1] Corresponding Author: Alberto Olivares-Alarcos; E-mail: alberto.olivares.alarcos@gmail.com

[2] Along the paper we will refer to distances and similarities equally, in fact mapping between the two measures is straightforward.

[3] For instance, *SNOMED-CT*, https://www.snomed.org/snomed-ct, one of the biggest terminologies of the medical domain, is organized as a taxonomy having 18 top categories, being one of them, *Pharmaceutical/biologic products* roughly corresponding to *Drug*.

[4] https://www.drugbank.ca/

- Two drugs are similar if their chemical compositions are similar.

With these intuitions in mind we can consider three basic types of distances: 1) Textual (string-based), 2) Semantic, and 3) Chemical (Molecular).

Textual similarities are based on computing some overlap between the two strings (at the level of words, n-grams, lemmas, etc.) and include distances as *Damerau-Levenshtein (string edit)*, *Dice*, *Jaccard*, and many others. See [2] for a survey. Semantic measures are computed within a semantic space, usually an ontology or a knowledge base. The most popular are based on WordNet[5] on which several distances can be computed based on the best path between the corresponding *synsets*[6], [3]. Different approaches differ on the way of computing the paths (type of relation, weighting, etc.). From them we have chosen *Leacock and Chodorow* metric [4] computed as shown in Eq. (1) where $length$ is the length of the shortest path between the two concepts (using node-counting) and D is the maximum depth of the taxonomy.

$$(1)\ Sim(d1,d2) = -\log\left(\frac{length}{2D}\right)$$

Similar approaches have been followed in [5] over the *Wikipedia* and, Within the medical domain, in [6] 0ver *UMLS*[7].

Chemical (Molecular) measures try to compare the 2D, representations of the molecules using the fingerprints vectors[8].

Sometimes units to be compared can be represented or mapped into a vectorial space. In this case these units are vectors in a n dimensional space. The usual way of defining distances is as instantiations of the *Minkowski* family. *Minkowski* distances are induced by l_p norms, see Eq. (2). For p = 1 we get the popular l_1 norm, i.e. Manhattan distance, Eq. (3), and For p = 2 the ordinary Euclidean distance, l_2 norm, Eq. (4)

$$(2)\ D_p(\vec{x},\vec{y}) = \|\vec{x}-\vec{y}\|_p = \sum_{i=1}^{d}(|x_i-y_i|^p)^{1/p}$$

$$(3)\ D_{11}(\vec{x},\vec{y}) = \sum_{i=1}^{N}|x_i-y_i|$$

$$(4)\ D_{12}(\vec{x},\vec{y}) = |\vec{x}-\vec{y}| = \sqrt{\sum_{i=1}^{N}|x_i-y_i|}$$

One issue of these formulations is that it is implicitly assumed that the dimensions are equally important. An alternative is weighting the dimensions, leading to the *Mahalanobis metric*, Eq. (5), being M a square semidefinite positive $n \times n$ matrix[9]. When M is the identity *Mahalanobis metric* reduces to *Euclidean Distance*. When M is a diagonal

[5] https://wordnet.princeton.edu/

[6] http://wn-similarity.sourceforge.net/

[7] https://www.nlm.nih.gov/research/umls/

[8] A fingerprint is a vector, each element of which describes the presence of one or more substructures in a molecule, with typical fingerprints containing a few hundred or a few thousand elements, and with two molecules being considered to be similar if their fingerprints share common values for many of the constituent elements.

[9] In the original formulation of *Mahalanobis*, M was the inverse of the covariance matrix of the input vectors.

matrix the components are weighted by the value of the corresponding entry in the diagonal. In the general case M captures correlations between features implicitly existing in the dataset.

(5) $D_M(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T M(\vec{x}-\vec{y})}$

2. Related Work

Many applications in the medical domain are heavily based on the use of distance measures between medical entities, frequently quite simple. Our approach can result on improvements on these applications. [7] use distance measures for Question Answering. [8] survey the task of Finding Patterns in Annotation graphs. The group lead by M.E. Vidal, at Simon Bolivar University, has very nice contributions to this area, see [9]. Other related applications are drug Discovery [10], drug targets [11] and drug interaction [12]. [13], perhaps the closest work to ours, study drug similarities including Chemical-based, Ligand-based, Expression-based, Side-effect-based, and Annotation-based.

3. DrugBank

The data used for the experiments come from the database *DrugBank*, a unique bioinformatics/cheminformatics resource combining detailed drug (i.e. chemical) data with comprehensive drug target (i.e. protein) information [1]. It is the most complete database about drugs which there exists nowadays.

The latest release of *DrugBank* contains 11,002 drug entries including 2,503 approved small molecule drugs, 943 approved biotech (protein/peptide) drugs, 109 nutraceuticals and over 5,110 experimental drugs. Additionally, 4,910 non-redundant protein (i.e. drug target/enzyme/transporter/carrier) sequences are linked to these drug entries. Each DrugCard entry contains more than 200 data fields with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data. Some of those fields are textual, like the ones used in the text based similarity explained in the sub-section 4.1. Some other fields are related to the chemical structure of a drug and have been used in the experiment explained in sub-section 4.3.

4. Computing Similarities Between Drugs

The implementation of the three similarities can be found on a free access repository on GitHub[10]. The evaluation framework is detailed in Section 4.4.

4.1. Textual Similarity

Text similarity is the task of determining the degree of similarity between two texts. Texts length can vary from single words to paragraphs or even complete documents. In our case, the texts are a concatenation of different textual fields extracted from the *DrugBank*

[10] https://github.com/albertoOA/Medical-Entities-Similarity-Measurements

database. Since the way of computing the text-based similarity relies on the bag of words (BoW) paradigm, simple concatenation of textual fields seems to be a good choice.

The computation of text similarity is a very difficult task for machines. This is mainly due to the enormous variability in natural language, in which texts can be constructed using different lexical and syntactic constructions. Even so, computing text similarity has been for several years a fundamental means for many NLP tasks and applications. [14,15,16].

Our aim is to find a measure of similarity among the drugs found in *DrugBank* by means of text similarity. To this purpose, the drugs were represented in a vector space model, specifically, the data fields: description, indication and pharmacodynamics all expressed in natural language were concatenated and, after removing stop words and transforming to lowercase, their term frequency-inverse document frequency (tf-idf) representation was computed. In this case, each document used to compute the tf-idf is the concatenation of the textual fields of each drug, while the corpus is formed by all those documents as a whole. Thus, the data is represented as the matrix $M \varepsilon \mathbb{R}^{n \times d}$, where n is the number of drugs and d the number of words in the whole corpus.

Usually, the number of terms within a corpus is large, this together with the fact that only few terms appear in a specific document give room to a sparse matrix. The high dimensionality and sparseness of the matrix M entail to a well-known phenomenon called 'curse of dimensionality'. In a nutshell, we lose statistical significance and the Euclidean distance becomes meaningless.

Therefore, we perform a dimension reduction of the vector space model we have computed. Specifically, we use Latent Semantic Indexing (*LSI*), also named Latent Semantic Analysis (*LSA*). LSA uses *SVD* to rank the most discriminative and useful information from our data. The distance matrix is computed using the Euclidean distance over the dimensionally reduced data. Of course, some information is lost after the reduction. The similarity is obtained ($similarity = 1 - distance$) after having normalized the distance matrix from zero to one.

4.2. Taxonomic Similarity

DrugBank contains two kinds of taxonomic structures: a set of fields named 'Classification' and the ATC Codes[11]. For this project, we have chosen to use the first one (Classification) to build a graph which is used to compute the similarity among the drugs. The second graph (ATC Codes) is used to evaluate the result. The classification field of DrugBank has 5 levels in total, enumerated from the highest to the lowest:

- Kingdom - Organic or Inorganic
- SuperClass - e.g. Organic Acids
- Class - e.g. Carboxylic Acids and Derivatives
- SubClass - e.g. Amino Acids, Peptides, and Analogues
- DirectParent - e.g. Peptides (can coincide with SubClass)

[11]The Anatomical Therapeutic Chemical (ATC) Classification System is used for the classification of active ingredients of drugs according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties.

The semantics of our taxonomy has only one sort of relationship: 'is-a' relationship, (e.g. Acetaminophen is-a SubClass of Benzenoids, or which is the same, Benzenoids is-a SuperClass of Acetaminophen).

We have used the classification tag in the DrugBank database to construct 2 trees (one for each *Kingdom)* of drugs) of 6 levels (depth equals to 5) which would connect the drugs in the database through undirected edges. Two different cases were contemplated: unweighted and weighted graphs.

We compute the distance between every pair of drugs as the length of the shortest path between them for two different cases: unweighted and weighted edges. As said in Section 1, we have used *Leacock and Chodorow* measure, Eq. (1).

4.3. Molecular Similarity

In principle, molecules that are structurally similar are likely to have similar properties. Thus, measures of structural similarity play an important role in cheminformatics for applications such as similarity searching, database clustering and molecular diversity analysis.

The main elements of any similarity measure based on Molecular Structure are:

- **Representation or Descriptor.** It is used to characterize the two molecules that are being compared. Among all the possible descriptors we use the fingerprints.
- **Weighting Scheme.** It is used to reflect the relative importance of different parts of the representation. No weights are used in this project.
- **Similarity Coefficient.** It is used to quantify the degree of resemblance between two appropriately weighted structural representations. In our case, we use the Tanimoto (Jaccard) Coefficient.

In our approach, we first calculate the fingerprints of each drug, using the information about the Molecular Structure which DrugBank contains. Although molecular description can be obtained in two or three dimensions, we used 2D fingerprints since the number of drugs with 3D information is limited in DrugBank and actually, even though it does make a different, there is not any instance of 3D representation as well established as the fingerprints in the case of 2D representations [17]. In this work, two of the most well-known types of fingerprints: MACCS and ECFPs have been used. Using the fingerprints, we compute the similarity among all of them using the Tanimoto Coefficient. The computation of the Tanimoto Coefficient for two binary vectors (a and b) of length k is defined as:

$$\frac{\sum_{j=1}^{k} a_j \times b_j}{(\sum_{j=1}^{k} a_j^2 + \sum_{j=1}^{k} b_j^2 - \sum_{j=1}^{k} a_j \times b_j)} \tag{1}$$

4.4. Evaluation

In order to evaluate the goodness of the three approaches to compute distance measures between drugs, we have performed two different evaluations: Clustering and Ground Truth based.

4.4.1. Clustering

Similarity values were used to cluster the drugs into groups using clustering. The performance of the clustering is analyzed comparing the obtained clusters against classification of drugs which serves as reference and is provided by the well known ATC Code Classification.

Drugs were clustered using Spectral Clustering, which is based on the use of the spectrum (eigenvalues) of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions. One drawback or limitation of this sort of clustering is that it is necessary to know in advance the number of clusters.

The Anatomical Therapeutic Chemical (ATC) Classification System is used for the classification of active ingredients of drugs according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties.

Each drug has associated at least one ATC Code which classifies all the drugs into different groups. Every ATC Code includes 7 alphanumeric characters which represent 5 different levels into the taxonomy or classification. This is to say, the first level can be understood as a subclass of 'Drug', which divides the drugs into 14 groups, the second level includes the subclasses of all the 14 previous groups, the same applies for the next levels. For instance, we analyze the ATC Code of *'Furosemide'* (C03CA01) level by level:

- First level: It indicates the anatomical main group and consists of one letter, there are 14 main groups, e.g. ***C*** Cardiovascular System
- Second level: the therapeutic subgroup and consists of two digits, e.g. ***C03*** Diuretics.
- Third level: the therapeutic/pharmacological subgroup and consists of one letter, e.g. *C03**C*** High-ceiling diuretics.
- Fourth level: the chemical/therapeutic/pharmacological subgroup and consists of one letter, e.g. *C03C**A*** Sulfonamides.
- Fifth level: the chemical substance and consists of two digits, e.g. *C03CA**01*** Furosemide.

For simplicity, we have decided to use fourteen as the number of clusters, the same number of categories of the first level of the ATC Code.

4.4.2. Ground Truth

The external (direct) evaluation consisted of comparing the computed similarities values with the degree of similarity between 100 pairs of drugs which were annotated by experts. That annotated data has been taken from [18]. Specifically, the ground truth was built using the opinion of 143 experts, who provided Yes/No decisions on a set of 100 DrugBank V3.0 molecule-pairs. Experts were asked to answer with Yes/No to the question: Are this pair of molecules similar?. The answers were collected and a distribution of Yes/No answers was computed. In this project, we use the proportion (percentage) of 'Yes' answers as degree of similarity. Of course, the reader should note that the experts were not asked about the degree of similarity.

In order to evaluate how our similarity measures are related to the ground truth values, we have studied three different dimensions:

- **Order.** We order the pairs by the value of their similarity in both cases, the list annotated by the experts and the one with our similarity measures. The correlation between both ordered lists is studied using Kendall's Tau Correlation.
- **Value.** The correlation between the value of the two lists (ground truth and computed in this project) is studied using Pearson's Correlation.
- **Threshold.** We have selected a threshold to classify the pairs of drugs into two different categories: similar and non-similar. If their similarity value is greater than the threshold, then, the drugs are similar. The threshold we have chosen is 0.85. The reason is because one of our similarity measures, the Tanimoto Coefficient, is considered relevant from that value. Then, we compute the precision and the recall of the classification process.

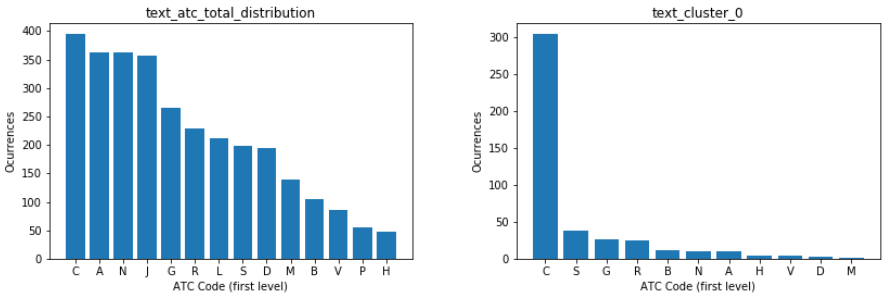
5. Experiments

Three different similarity measures between drugs have been implemented using Python and computed over DrugBank. The evaluation of all these measures have been done following the process explained in Section 4.4. All of this work was done inside of the framework of a master thesis[19], which could be read for more detailed information about the obtained results. In this paper, we just show the result of two similarity measures, which correspond to the best performance in each of the two evaluations. On the one hand, the best results for the clustering has been obtained for the textual based similarity. On the other hand, the best performance of our similarities against the ground truth appeared during the molecular structure based similarity.

5.1. Evaluation results: Clustering

Always that it is available a reference cluster to compare with, the performance of a clustering can be studied using the measure of *Purity*. In our case, we have that cluster to compare with (ATC Code). However, the results are not really good so we decided not analyze them using the value of *Purity* but just do it visually, using the distribution of the ATC Code of the clustered drugs showed in the histogram of each cluster. This analysis is divided into three different groups of obtained clusters:

- Clusters in which the most common ATC Code represents a good percentage of the total number of occurrences of the ATC Code within the complete set of used data. This could be equivalent to the notion of *Purity*, meaning, how good we are grouping in the same cluster all instances of drugs with a certain ATC Code. There are some clusters which show this behavior though, the best instance is shown in Figure 1. As we see, from the total number of instances of the ATC Code *'C'* (a), we have been able to capture around the 75% in the cluster 0 (b). Thus, our similarity measures can be considered as good.
- Clusters in which the most common ATC Code appears clearly more times within the cluster (clearly predominant) than the rest of ATC Codes included in the cluster. Even though it is not exactly the same, this is somehow related to the notion of *Purity*. There are several clusters which show this characteristics, however, here we show one of the examples (see Figure 2). As we can see, in this cluster (number 3), the most predominant ATC Code, *'A'*, is clearly predominant over the rest



(a) Total Distribution of ATC Codes. (b) Distribution of Cluster 0.

Figure 1. Total real distribution of the ATC Codes for the drugs used in the text similarity evaluation (a) and distribution of one of the obtained clusters (b).

.

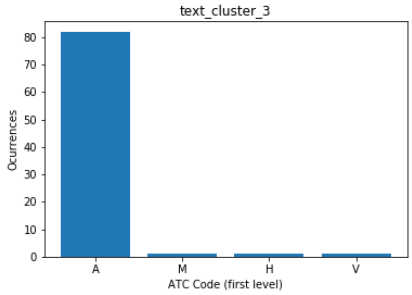


Figure 2. Cluster obtained during the textual based experiment. ATC Code 'A' is clearly predominant over the rest.

of ATC Codes, so we can say that this cluster corresponds to a cluster with drugs with ATC Code *'A'*. Thus, our similarity measures are good.

- Clusters which are a bit meaningless for us because either they cannot be included in one of the previous cases (e.g. there is not a clear predominant ATC Code within the cluster) or because the number of drugs within the cluster is too small.

5.2. Evaluation results: Ground Truth

As stated before, we have computed two different similarity matrices for the Molecular Structure based experiment: one with MACCS fingerprints and another one with ECFP fingerprints. In the Table 1, the results for this evaluation are shown.

The best result is clearly the case in which MACCS fingerprints are used, since the values are always better. Two comments are worth for this experiment. On the one hand, the result is quite better than in the case of using textual and taxonomic information because of two reasons: we used as threshold 0.85, the value Tanimoto Coefficient (similarity used here) is relevant from and because the way in which the ground truth was built (experts used the molecular structure of the drugs). On the other hand, we can claim that there is not correlation between the inferred rank by using our similarity and the ground

Sort of Fingerprint	ECFP	MACCS
Pairs in ground truth	97	97
Pairs in computed similarity	96	96
Kendall's τ	-0.0404	0.0601
Pearson's Correlation	0.8886	0.9186
Accuracy	0.7708	0.8854
Recall	0.12	0.76

Table 1. Direct Evaluation against a ground truth of the Molecular Based Similarity

truth, this is true for the three experiments. A possible reason is that experts were not asked about which degree of similarity have those drugs, neither they were asked to say how similar are two drugs in comparison to another two other ones.

6. Conclusions

Distance measures between medical entities are known to be important for many NLP tasks. Computing distances between drugs is specially challenging due to the different facets implied in different tasks. For instance, in text mining applications, textual approaches seem to be the most appropriate, for more semantic-based applications, as QA or Linguistic Inference, taxonomic approaches seem to be better, while for drug reposition molecular approaches are the most likely. Three different similarity measurement over drugs from DrugBank have been implemented: textual, taxonomic and molecular. To our knowledge there is no other work which includes these three measures within the same framework. A evaluation of the implemented similarities has been performed by means of both indirect (Clustering) and direct (Ground Truth) evaluation. All the implemented code is shared as open source code under MIT License on GitHub, which leads to the main contribution of this work, since our similarities could be used by other researchers to perform tasks as the ones proposed in Section 2.

The Clustering evaluation has provided lights and shadows, while in some cases we have been able to cluster properly the drugs based on their ATC Codes, we have not in several cases. This does not strongly implies our similarity measures are not good. Spectral Clustering, used in this work, and graph-based semi-supervised learning algorithms, in general, are well known to be sensitive to how graphs are constructed from data. In particular if the data has proximal and unbalanced clusters these algorithms can lead to poor performance.

On the other hand, some promising results have been found in the evaluation based on the ground truth, specially, for the similarity based on Molecular Structure. Nevertheless, the results are not definitive, a need for a larger ground truth is clear. Of course, we have not found a larger one on the literature, so the only solution would be to build one, which implies the help of experts. We also claim that there is not correlation between the inferred rank by using our similarity and the ground truth, this is true for the three experiments. A possible reason is that experts were not asked about which degree of similarity have those drugs, neither they were asked to say how similar are two drugs in comparison to another two other ones.

Considering the evaluation process, we claim that there is still work to do on improving our similarity measures so that they could be more valuable for the research

community so that they could become widely used. However, this work has shown we are going in the right direction.

Acknowledgements

We are grateful to the comments and suggestions from two anonymous reviewers. This work was partially supported by project GRAPHMED (TIN2016-77820-C3-3R).

References

[1] *Wishart D. S., Feunang Y. D., Guo A. C., Lo E. J., Marcu A.,; Grant J. R., Sajed T., Johnson D., Li C., Sayeeda Z. (2018)*: DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic acids research, vol. 46, 2017, pages 1,074-1,082, Oxford University Press.

[2] *Gomaa G., Fahmy A. (2013)* A Survey of Text Similarity Approaches, International Journal of Computer Applications.

[3] *Pedersen T., Patwardhan S., Michelizzi J. (2004)* WordNet::Similarity - Measuring the Relatedness of Concepts. Proceedings of the 19th National Conf. on Artificial Intelligence (AAAI-04), pp. 1024-1025.

[4] *Leacock C., Chodorow M. (1998)*. Combining local context and WordNet similarity for word sense identification. In WordNet: An electronic lexical database, vol. 49, number 2, pages 265-283

[5] *Gabrilovich E., Markovitch S. (2007)* Computing semantic relatedness using Wikipedia-based Explicit Semantic Analysis. 20th Int Joint Conference on Artificial Intelligence (IJCA) p. 16061611.

[6] *McInnes B., Pedersen T., Pakhomov S. (2009)*. UMLS-Interface and UMLS-Similarity: Open Source Software for Measuring Paths and Semantic Similarity. AMIA Annual Symposium. Pages 431-435.

[7] *Goodwin TR., and Harabagiu S. (2016)* Medical question answering for clinical decision support. 25th ACM International on Conference on Information and Knowledge Management, pages 297-306.

[8] *Ding H., Takigawa I., Mamitsuka H., Zhu S. (2013)* Similarity-based machine learning methods for predicting drug-target interactions: A brief review. Briefings in Bioinformatics, 2013.

[9] *Palma G., Vidal ME, Raschid L. (2014)* Drug-Target Interaction Prediction Using Semantic Similarity and Edge Partitioning. Semantic Web Conference 2014, pages 131-146

[10] *Lipinski CA:, Lombardo F., Dominy BW., Feeney PJ. (2012)* Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced drug delivery reviews, 64:4.

[11] *Overington JP. Al-Lazikani B., Hopkins AL. (2006)* How many drug targets are there?. Nature reviews Drug discovery, 5(12), pages 993-996, 2006.

[12] *Melnikov MP., Vorobkalov PN. (2014)* Retrieval of drug-drug interactions information from biomedical texts: Use of tf-idf for classification. In Knowledge-Based Software Engineering, pages 593-602, 2014.

[13] *Perlman L., Gottlieb A., Atias N., Ruppin E., Sharan R. (2011)* Combining drug and gene similarity measures for drug-target elucidation. Journal of Computational Biology, 18(2), pages 133145, 2011.

[14] *Kenter T., De Rijke M. (2015)*. Short text similarity with word embeddings. Proceedings of the 24th ACM international on conference on information and knowledge management. Pages 1411-1420.

[15] *Kashyap A., Han L. Yus, R., Sleeman J., Satyapanich T., Gandhi S., Finin T. (2016)*. Robust semantic text similarity using LSA, machine learning, and linguistic resources. Language Resources and Evaluation Journal (Springer). Pages 125-161.

[16] *Ho PH., and Nguyen NAT., Vo, TH. (2018)*. DNA Sequences Representation Derived from Discrete Wavelet Transformation for Text Similarity Recognition. Modern Approaches for Intelligent Information and Database Systems (Book, Springer). Pages 75-85.

[17] *Willett P. (2014)*. The calculation of molecular structural similarity: principles and practice. Molecular informatics (Wiley Online Library), volume 33, pages 403-413.

[18] *Franco P., Porta N., Holliday JD., Willett P. (2014)*. The use of 2D fingerprint methods to support the assessment of structural similarity in orphan drug legislation. Journal of Cheminformatics, vol. 6.

[19] *Olivares-Alarcos A. (2018)*. Semantic Distances between Medical Entities. Master Thesis, Universitat Politecnica de Catalunya.