A Family of Algorithms for Patient Similarity Based on Electronic Health Records

Yang Liu

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Computer Science at

Concordia University

Montréal, Québec, Canada

September 2022

Concordia University

School of Graduate Studies

This is to certify that the thesis prepared

By: **Yang Liu**

Entitled: **A Family of Algorithms for Patient Similarity Based on Electronic Health Records**

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

______________________________ Chair
Dr. Nematollaah Shiri

______________________________ Examiner
Dr. Olga Ormandjieva

______________________________ Examiner
Dr. Nematollaah Shiri

______________________________ Supervisor
Dr. Vangalur Alagar

Approved by ____
Leila Kosseim, Chair
Department of Computer Science and Software Engineering

________________ 09\08\2022 ____________________________
Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

A Family of Algorithms for Patient Similarity Based on Electronic Health Records

Yang Liu

Patient similarity is an emerging field of study facilitating health care analytic of big data pertaining to patients. Its major goal is to rank or cluster patients so that each cluster exhibits one aspect of tightly related patient characteristic. These characteristics include diseases, drugs take, risk factors, life styles, habits and ethical aspects. The rapid adaption of Electronic Health Record (EHR) in hospitals and other governmental heath care organizations to store such a variety of patient information provides a comprehensive source for efficient health care delivery, for data-based analytic of patient-centric individualized perspective prediction, and decision making. It is in this context that the thesis is making contributions on structuring an EHR as a vector of multifaceted components, where each component may be an aggregation of sub-components and each sub-component has a strict type. The operations on each sub-component are part of the typing scheme, and permits semantic-based similarity assessment on each component. The suggested EHR structure is both generic and extendable. The scoring functions that measure the similarity between pairs of components are rigorously defined with respect to domain semantics and user semantics. The weighted average of the scores of components, where the weights are part of user semantics, calculates the similarity between records under analysis. Several examples are shown to comprehensively explain the behaviour of functions. Drug-Drug similarity, and patient-patient similarity analysis based on it are discussed. Experimental results are given and their merits are explained.

Acknowledgments

Completing this thesis was not an easy task, I would not be able to finish without the help from all sources, I would give my thanks to the people who helped me along the way.

First and foremost, I would like to give my great gratitude and respect to my supervisor Dr. Vangalur Alagar, for always being tolerant and informative. His knowledge and guidance have helped me throughout my Master's study. Dr. Alagar always teaches me with a great amount of concern and shapes my academic mindset with precise and clear explanations, even during the pandemic. There are no words that I can express my appreciation for him. Also, there's great thanks to the seniors Alaa Alsaig and Ammar Alsaig.

I would also appreciate the help from the Gina Cody School of Engineering and Computer Science at Concordia University, thanks to Dr. Aiman Hanna, Ms. Halina Monkiewicz and Ms. Samantha Singh and other faculty members at Concordia University and the Webster Library for being tolerant and warmhearted.

Thank to my girlfriend Xueli Chen for all her precious love and warm support. Meeting her is the luckiest thing that ever happened in my life.

I would thank my friends, Eddie Zhou, David Liu, Wei Qi, Tianyi Qi and others who bring me a lot of knowledge and joy and let me know I am not alone in this world.

Let me express my thanks to my parents for their concerns from home and for always supporting and encouraging me to do what I wanted to do. Their open minds have given me so much potential to complete all the things.

Thanks to the all-knowing god and the almighty wisdom he gave us to make advances in personal and academic life.

Last, let me express my sincere gratitude to all the people who are with me along the journey, and who helped me during the tough days, thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Hospitals and medical clinics all over the world have long been maintaining medical records of patients. Such a paper-based record, called *Electronic Patient Records* (EPR), contained patient medical charts, medical history, medications, and allergies of a patient. This EPR documentation was regarded as authentic and shared only by a small group of physicians and nurses. So, the hospital administration used to maintain a separate database, and service providers had their own database. Since the early 2000, the term *Electronic Medical Record* (EMR) was used (Chang & Gupta, 2015) to refer to a combination of EPR and hand written notes. They were used only by medical treatment providers for diagnosis and treatment of the patient. In 2005, government organizations and healthcare providers realized the need to eliminate error-prone handwriting notes, and consolidate health information included in EMR with a lot more patient data such as vital signs, clinical visits, lab data, medications, and imaging reports into a digital record (Gunther & Terry, 2005). After the advent of digital society and internet, hospitals and clinics started a systematic recording of this extended set of healthcare information in digital form. This gave birth to *Electronic Health Record* (EHR). In USA, Australia, and Canada (Chang & Gupta, 2015; Gunther & Terry, 2005; Hecht, 2019; Watzlaf, Rhia, Zeng, Jarymowycz, & Firouzum, 2004) EHR adoption rates in hospitals steadily increased, once they recognized the potential advantages of EHR which include *a unified repository of patient records*, *distributed sharing of information*, *efficient communication among EHR users*, and *a longitudinal health record that can be developed effectively.* Thus, EHR concept was originally conceived to facilitate only healthcare delivery.

Of late, many researchers (Birkhead, Klompas, & Shah, 2015; Kruse, Stein, Thomas, & Kaur, 2018; S. Lee et al., 2017) have recognized that EHR databases are "potentially an optimal data source for research", and they can help to *support public health safety*, *promote public health surveillance*, and *serve as valuable data source for healthcare research.* There are many challenges to achieving these lofty goals (Hecht, 2019). The three major challenges brought out by Lee (S. Lee et al., 2017) are data structuring, data integration, and combining structured and unstructured data for data mining and ML-based deep analysis. It is in this context, we set the goal of this thesis to investigate similarity analysis methods on EHRs restricted to structured data.

The research gap that exists currently between "EHR data structuring" and "EHR similarity analysis" has been well documented in (Pokharel, 2020). Primarily focusing on temporal data, this work has proposed different kinds of hierarchical data structures and algorithms based on these structures to assess similarity of clinical data in patient records. Such kind of data, mainly heterogenous and sequential in nature with time stamps, has also been studied by other researchers (Z. Huang, Dong, Duan, & Li, 2013; Kunimoto1, Vogt1, & Bajorath1, 2016; J. Lee, Maslove, & Dubin, 2015) for similarity assessment without resorting to hierarchical data structures. However, these works do not emphasize types of attributes for recording information. As a consequence, methods used by them cannot be checked for correctness and validity across different datasets. In this thesis, EHR similarity analysis are tied to use only operations defined on types of attributes used in EHR structure. That is, there is no gap between EHR structuring and EHR similarity analysis. In fact, our approach allows integrating the domain-level semantics of attributes with user-level semantics on "match making" between EHRs and analyst query.

1.1 Characteristics of EHR Data

Every patient has a *uniquely identifiable* EHR which contains the health information of that patient. Now, there is a general consensus (S. Lee et al., 2017) on the following list (non-exhaustive) of information to be included in the EHR of a patient.

- Patient demographics
- Vital signs

- Medical history (Diseases and types, treatment times, clinical visits)
- Diagnosis (date, progress notes, drugs administered)
- Medication List (for each disease type, dosage)
- Lab and test results
- Radiology images

But for data on “Radiology Image” and long textual information on “progress notes” the rest of the information in the above list can be put in a structured format wherein each structured item is either *atomic* or *compound*. In this thesis we do not consider inane data and unstructured (mainly long texts) data. In the above list, first name, last name, age, marital status, and gender are atomic, and the rest are compound. An example of compound data in the above list is “patient demographic data”. It will include many atomic pieces, such as first name, last name, gender, marital status, and compound data, such as date of birth, and primary care provider. Each one them can be assigned a type, as we explain in Chapter 2. A review of the current EHRs (Chang & Gupta, 2015; Gunther & Terry, 2005; S. Lee et al., 2017) and healthcare datasets from hospitals (Janosi, Steinbrunn, Pfisterer, & Detrano, 2017; Miriam Seoane Santos, 2017) reveals that currently the above list of information is recorded in EHRs in many different formatsn. Some, but not all, EHR data sets specify the types of attributes and the unit of measurements (for vital signs). Missing types, and incompleteness in data values make healthcare delivery and research hard. With strict type information and avoiding incompleteness in attribute values, the EHR datasets will have the following characteristics:

- *Generality*: An attribute, whether its domain has atomic or compounded value, will have a type. Compounded attributes will convey comprehensively the relationship among compounded elements. The EHR structure will be extendable to comprehensively include as many typed attributes as are necessary to make the structure general enough for many types of analyses. Of course, in practice it is not possible to obtain “model completeness” because of the growing futuristic demands in health domain. However, the goal is to make the structure general and flexible (extendable) so that the *projection* of the EHR on a subset of its attributes will suffice for an analysis context. It should be possible for experts to decide the choice of attributes and their types, so that it is possible to combine the analyses of results of projected EHRs without ambiguities.

- *Usability*: The recorded information in a EHR will be precise in order to be understood and used without ambiguity. Only attributes of the same type will be comparable and their values manipulable using the operations defined for the type of that attribute. Hence, similarity assessment between pairs of attributes is meaningful.

- *Shareability*: For effective healthcare delivery, the fulfillment of "usability" criterion allows information in EHRs to be shared and interpreted correctly by doctors, nurses, researchers, administrators and other groups of authenticated users. Meanwhile, the shared information can be regulated among different types of users, respecting access control policy governing users and patients.

- *Patient-centric*: A patient-centric healthcare model can provide additional attributes that have an impact on each patient. Type system will assign types to such attributes. A patient may be given the privilege to introduce and control certain private information at the attribute level, which will be respected by the similarity calculation algorithm.

1.2 Research Contribution

Viewing a EHR as a "vector" of *attributes* (features), where each attribute is either an atomic type or compound type, in this thesis we investigate type-strict methods for comparing pairs of attributes across records and assess their "closeness". Conceptually, it is a vector whose elements may have different types and hence structures in an implementation. The thesis proposes a user-centric approach towards EHR analysis. That is, both domain-level and user-level semantics are integrated in assessing closeness of attributes in records using the operations that are specific to types. Domain level semantics is often supported by Ontology for concept terms in healthcare. User-level semantics is gathered by allowing a research analyst to specify the options for attribute matching and preference semantics for assessing the score of closeness for a pair of attributes. However, when all EHR records are assessed pairwise for similarity, all EHR records are treated as “equal citizens”. We construct functions to calculate similarity scores between pairs of attributes of the same type, and use these scores to assess the similarity between two records or between the user “query” and EHR records. This user-centric similarity assessment is a new contribution to healthcare analysis. No other researcher has proposed a user-centric similarity calculation for health records. Below we explain how the rest of the thesis is structured. In every chapter we compare our approach and methods with other related work, and highlight our new contribution.

Basic concepts on “Attributes and Types” are given in Chapter 2. The goal is to make the thesis self-contained with respect to EHR structure on which EHR analysis is defined using type-specific operations. Also, this exposition makes clear the basic types (such as Numeric, String, Categorical, Nominal, Enumerated, Range), and higher-order (compound) types (set, Record, Tree, Bag) used for EHR model.

In Chapter 3, a few hospital clinical records are reviewed and compared to illustrate the most commonly occurring attributes (and their types) in these records. We define the EHR structure as a “vector”, an ordered collection of different types of attributes. Our EHR type is not relational. It can accommodate temporal values associated with clinical data, sequence type for recording texts coded as sequences (strings), and set types for recording sets of numeric or categorical or textual keywords. By combining set and sequence types in EHR structure we can record longitudinal data, which is data collected sequentially from the same respondents over time.

In Chapter 4 we review many similarity functions proposed in the literature, compare those that have been used in healthcare, and bring out the flaws and inadequacies in most of them. We select two functions that fulfills our criteria and based on them we build new similarity assessment functions. We evaluate the behavior of our similarity functions theoretically, and study their relative efficiency by observing experimental results. These functions are used in assessing drug-drug similarity and patient-patient similarity in later chapters.

The user-level semantic options, the structure of analyst query, and a general algorithm to compute and rank a database of EHR records (their projections) are given Chapter 5. All scoring functions are constructed from the functions proposed in Chapter 4.

In Chapter 6 we discuss the attributes selected to model a drug, and emphasize their significance in similarity assessment of drug records. We use the scoring functions and the algorithm discussed in Chapter 5 to assess the pairwise similarity of drug records, and also rank drug records against a analyst query. The method is implemented. From results on a few examples and case study we validate the merits of our approach.

Similarity between patient records can be assessed in many ways (Chan, Chan, Cheng, & Mak, 2010; Ferdousi, Safdari, & Omidi, 2017; Girardi, Wartner, Halmerbauer, Ehrenmüller, & Kosorus, 2016; Harispe, Sánchez, Ranwez, Janaqi, & Montmain, 2014; Hu et al., 2017; J. Lee et al., 2015; Mabotuwana, Lee, & Cohen-Solal, 2013; Zhang, Wang, Hu, & Sorrentino, 2014). However, given the wide variety of attribute types that model a patient, we thought the similarity will be more meaningful when restricted to EHR projections on a subset of tightly related attributes. So, in Chapter 7 we study the similarity between a pair of patients induced by the similarity of drugs for a specific disease type. Based upon the outcome, we explain how different types of further analysis can be done. The results from a case study validates the merits of our approach.

The thesis is concluded in Chapter 8, with a summary of contributions, comments on the relative merits of our work when compared to other work, and suggestions on future extensions.

Chapter 2

Attributes and Types

In natural language the term *attribute* is used to convey the meaning of "something is attributed to someone or something" (Merriam Webster, 1828), as in the text "The doctor attributes the health problem to irregular diet". In data management and data analysis, an *attribute* is a *characteristic* (or *feature*) of an *entity* (or *object*) of interest that is measurable either quantitatively or qualitatively. In data management and data analysis, an entity of interest can be *EMPLOYEE* or *STUDENT* or *PATIENT* or *VEHICLE*. An entity is modeled using a finite set of attributes. As an example, some of the attributes of *EMPLOYEE* can be $Name$, *Emp_ID*, *Experience*, The values of some of these attributes vary dynamically, and are measurable either numerically or descriptively. The set of attributes of each entity in data analysis provides a model of that entity. The set of *operations* defined on an attributes enables manipulating the data values associated with that attribute. The entity models together with operations on them make every entity a first class citizen in the data management system. In this thesis, our interest is in attributes that are used for modeling and analyzing EHR. Following the notion of types in programming and abstract data types (Dale & Walker, 1996), and guided by the fundamental principles on the quantitative approach for mathematical rezoning (Bennett & Briggs, 2015), in this chapter we give a brief account of basic *attribute types* and operations on them. We explain how using these results compound attributes and operations on them can be constructed.

2.1 Importance of Attribute Types

In order to enable data analysts choose relevant data sets for analysis, and systems designers to develop suitable algorithms for manipulating data for analysis it is necessary to formally define the *types* of attributes. A type, as understood in programming language design, is a *set* S of values together with a *set* O of operations on the set. For all simple data types, *equality* operation is defined in a natural manner. If the data type is *composed* from one or more data types, equality is defined on each simple data type in the composition. For any data type which allows arithmetic operations the set S is closed with respect to the set of arithmetic operations. Many data types allow *relational* operations. In general, every data type definition must include a set of well-defined operations on the set of elements in the domain of the type. Following this convention, in this section we review the types of commonly arising basic attributes.

Operations associated with a type play a significant role in *type safe* manipulation of attribute variables. Attribute values may be *numbers* or *symbols*. The value of a *symbol* is itself, unless it's specific semantics is defined in the domain to which it belongs. So, the operations adhere to the semantics of attribute domains. With two examples, we emphasize the importance of type-specific (semantics-based) interpretation of attributes and operations are necessary for correct data analysis.

Example 1. *If the type of an attribute is integer, then its* unit *must be given to semantically interpret the number (value) assigned to the number. As an example, the attribute* Height *of a person may be given as* Real*, however its values can be measured either in feet/inches or in meters. With this "semantics" for attribute value definition, the operations on real values applied to* Height *can be meaningfully interpreted. For attribute* Age*, the value is usually given as a "sequence of digits". It is assumed that the integer value of this sequence denotes the number of years completed since birth. Although many researchers associate* Integer *type for* Age*, it is incorrect because all "integer operations" are not allowed on age values. So, the type of* Age *attribute is better be defined as an* enumerated type *(sub-type of integer type)* $1, 2, \cdots, 125$ *(assuming the maximum age recorded in the database is* 125*). With this definition, only comparison operation and "additions for statistical computations" are allowed on Age attribute.*

Table 2.1: SNOMED CT Example

Concept	**Value**
Gastric Ulcer	397825006
Stress Ulcer	415623008
Gloma	393564001
Excision-action	129304002
Laser Device	122456005

Example 2. *An "integer" coding for certain concepts in healthcare domain must be treated differently from "integer value". In* Systematized Nomenclature of Medicine—Clinical Terms *(SNOMED CT) (El-Sappagh, Franda, & et al., 2018), a comprehensive medical terminology is used for standardizing the storage, retrieval, and exchange of electronic health data. In Table 2.1 we give the ontology values for some SNOMED CT "concepts". These values, although have "integer coding", are not integer types because each value "is a place holder" for a concept. So, their type is "categorical" whose semantics is given by the Ontology. In addition to equality (inequality) operation, additional operations induced by the Ontology structure are permitted on the codes.*

In this thesis we assume that an ontology support is available to help the analysis in healthcare applications. Ontology provides the semantic support for defining the correct set of operations necessary for an analysis. This brief background explanation is given just to motivate why a formal definition of type is necessary for analyzing datasets.

2.2 Simple Attribute Types

An attribute may characterize either *quality* or *quantity* of data. So, the two kinds of simple attributes are *quantitative attributes* and *qualitative attributes.*

2.2.1 Quantitative Attributes

A *quantitative* attribute can take *numeric* values that are either *discrete* or *continuous.* The type of attribute *Age* of an employee in a company is an enumerated set and hence it is discrete. If the policy of the company is "the minimum and maximum ages of employment are respectively 18 and 65", then the values of *Age* attribute are from the finite set of integers

Table 2.2: Inpatient Stays in a Hospital

State	**Diagnosis**	**Age Group**	**Number of Stays**	**Monthly Rate of Stays**
Alabama	Liveborn	[30-35]	[360,000, 365,000K]	[1000,1500]
California	Heart Failure	[60-80]	[221,000,221300]	[650,690]
Alabama	Heart Failure	[60-80]	[221,000,221300]	[1025,1050]
Arkansas	Pneumonia	[50-65]	[74,000,74,200]	[26,27]
California	Cancer	[50-75]	[56,000,,56,100]	[517,620]

$\{18, 19, 3, \cdots, 65\}$. Because such a set includes all integer values in the range $[18, 65]$, the type of *Age* is regarded as *enumerated* type, whose values are written $18, \cdots, 65$. Integer type attribute allows all integer operations. However, for enumerated types only a restricted subset of integer operations may be allowed. Continuous attribute types take *real* (float) numbers as values. The attribute *Weight* of an object is of continuous type. As an example, it can take values either from a finite set $\{10.34, 11.78, 25.31, 56.93\}$ of real values, or from an infinite set of real values $\{x \mid 10 \leq x \leq 100\}$. The relational operations $\leq$, $\geq$, $=$, and $\neq$ are common to both discrete and continuous attributes. In healthcare domain, the frequently arising attributes such as *Pulse Rate*, *Wait Time*, and *Room Capacity* are discrete types, and the attributes *Weight of Patient*, *Wait of Patient*, *Blood Pressure*, and *Tumor Size* are continuous types.

We use the term *interval* to denote a dense set of real contiguous values, and the term *range* to denote a finite set of successive integers. It is sometimes necessary to consider *range* and *interval* as types of attributes. In general, hospitals and census bureau may not publish exact data, but provide only statistical summaries on patients and populations. Such summaries involve attributes of “range” or “interval” type. As an example, instead of publishing the exact number of cancer patients, a hospital might publish a *Count* of *range* type, which reveals a number in a certain range. Similarly, *Average* (mean), *Standard Deviation*, *Covariance*, and *Regression* may be reported in *Interval* types. Table 2.2 gives a sample of an annual publication of patient statistics from a hospital. The type of the first column attribute (and second column attribute) is *categorical* (see Section 2.2.2). The values in these columns have semantics but no specific ordering. That is, the rows can be permuted to convey the same interpretation for the table. The attribute type of third (and fourth column) is *range* because age and the number of patients who stayed can takes any discrete integer value within a set of successive integers. For example, the age of patients with "Cancer in California" can be any integer between 50 and 75, and the total number of stays for all patients is any integer in the range $56,000$ and $56,100$. The attribute type of fifth column is *interval*, because if the total number of patients in a month is M, and n_i is the number of days patent n_i stayed then the " average number of stays" is $\frac{\sum_{i=1}^{M} n_i}{M}$, a real number in the interval shown.

The standard operations used for *range* and *interval* types of data are $=$ (equality), $\in$ (set membership), and relational operators ($<$, $\leq$, $\geq$, $>$) that are used for real and integer values.

2.2.2 Qualitative Attributes

A quality attribute is one that *describes* a specific feature, but in general cannot be given numbers for measuring it. Attributes that describe eye color, hair color, postal code, patient identity, medicine code, reliability, safety, availability, and professional status of an employee are some typical examples of quality attributes. In genera, such attributes are of type *Categorical* that describe *categories or levels.* Categorical type can be further refined as *Nominal*, *Ordinal.* Nominal type describe categories that do not have a specific order. The values of Nominal attribute are *symbols* that denote names of things. These names are *enumerable* and hence are *discreet.* Nominal type is further refined into *categorical* and *ranked categorical* (or *ordinal*) types. In categorical type, the listed values do not imply an ordering. An example of categorical type is *Color* whose values are listed in *set notation*, as in $\{Brown, Red, Green, Black\}$. Another example of nominal type is *Marital Status* whose values can be $\{Single, Married, Divorced\}$. In healthcare domain the commonly arising nominal attributes are *Gene Code*, *Blood Type*, and *Medicine Code.* Only equality operation $=$ is defined for categorical attributes. However, as explained in Example 2 it is possible to define additional operations for categorical attributes when an ontology semantic support is given. In a later section on analysis, we will return to this necessity of defining operations for categorical attributes based on ontology support.

To order categories, without any semantic significance of order, we use *ranked categorical* attribute type. We use sequence notation to show the order of its elements. As an example,

$$< Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday >$$

represents the ranked categorical type *Weekday* whose events are sequentially ordered. To display the order) the notation *follows* ($\prec$) is used. That is,

$$Sunday \ \prec \ Monday \prec \ Tuesday \ \prec \ Wednesday \ \prec \ Thursday \ \prec \ Saturday.$$

In such ordering, there is no implicit meaning of “greater than” (higher) or “less than” (lower). Thus, equality ($=$) and *follows* ($\prec$) are the only two operations defined for *ranked categorical* attributes.

In some applications, it may be necessary to bring in the semantics of “higher (superior)/lower” for ranked categorical attributes. As an example, the ranked categorical type *Status* of faculty members in a university whose values are

$$< Lecturer, AssistantProfessor, AssociateProfessor, Professor >$$

suggests “greater than” (low status to high status) relationship in the listing. That is,

$$Lecturerer \ < \ AssistantProfessor \ < \ AssociateProfessor \ < \ Professor.$$

Even though we can order the elements from lowest to highest, the “spacing between the values” may not be the same across the levels of the variables. In some universities, $AssociateProfessor''statusisgivenafterthreeyearsofexperienceinthe$Assistant Professor” category, while in some universities the promotion to *Associate* status may include additional requirements. So, assigning numerical “scores” such as $1, 2, 3, 4$ to the categories will not lead to accurate analysis of data.

The *Ordinal Attribute* has values that have *semantically meaningful* interpretations and *ranking* (ordering). The values, listed as a sequence, show the importance of listed values but may not indicate the *relative* measure of importance. In the former sense it is different from a ranked nominal, although in the later sense it looks similar to it. The attribute

Grade is usually assigned a sequence of symbolic values

$$\langle A^+, A, A^-, B^+, B, B^-, C^+, C, C^-, D, F\rangle,$$

where each symbol is assigned a semantic interpretation in terms of *Intervals* of numerical values. That brings out not only the *monotonic decreasing values* of symbols but also the relative differences between any two symbols. Another example is *PayScale* attribute whose values can be *alpha numeric symbols*, wherein each symbol $E\#$ is associated with an interval $[X, Y]$, where $X < Y$ are positive integers indicating the range of salary that an employee with pay scale $E\#$. However, the difference between adjacent categories do not necessarily have the same meaning.

In data analysis, it is necessary that the variables (of different types) have specific *levels of measurement*. For example, it would not make sense to compute an average on *color of balls*. We cannot talk about average of a nominal variable, because there is no intrinsic ordering of the levels of the categories. However, for variables on ordinal type the average can be computed provided a numerical equivalent of the ranked values are definable. For example, it is possible to compute the Grade Point Average (GPA) of a student if a numerical equivalent to each grade symbol as shown in Table 2.3 is available.

Table 2.3: Letter Grade - Interval - Numerical Measure

Letter Grade	**Interval**	**Numerical Equivalent**
A+	[97-100]	4.2
A	[93-96]	4.0
A-	[90-92]	3.7
B+	[87-89]	3.3
B	[83-86]	3.0
B-	[80-82]	2.7
C+	[77-79]	2.3
C	[73-76]	2.0
C-	[70-72]	1.7
D+	[67-69]	1.3
D	[65-66]	1.0
E/F	< 65	0.0

Notice that the spacing between the grade levels is not uniform, yet the meaning of GPA based on such scheme is accepted in many colleges and universities. Sometimes we need to analyse data variables that are "in between" ordinal and numerical. For example, many questionnaires use a five-point Likert scale with values "strongly agree", "agree", "neutral", "disagree" and "strongly disagree" to collect user responses. In order to have unbiased analysis results we will assume that the intervals are equally spaced. With this assumption, we can assign the numbers $4, 3, 2, 1, 0$ as equivalent respectively to the values "strongly agree", "agree", "neutral", "disagree" and "strongly disagree".

The *Binary Attribute* has only two values, often denoted as $\{0, 1\}$, or $\{False, True\}$ or $\{Yes, No\}$ or $\{Fail, Pass\}$. For example, the attribute *Result* can be given binary type with values $\{Pass, Fail\}$, and the attribute *Diagnosis* can be a binary attribute with values $\{Positive, Negative\}$. Binary attribute may be viewed as ranked categorical with interpretations $Fail < Pass$, $Negative < Positive$, and $False < True$.

2.3 Higher-order Types

The higher order types that commonly arise for EHR data modeling are *List*, *Set*, *Sequence*, and *Record*. The data type *String* may be viewed as a sequence of "characters" (Char), where "Char" is a simple type. In this thesis, "string" is regarded as a simple type because it is a standard basic types in all programming languages. In our EHR modeling, we use Set, Sequence, and Record types. Informally, $Set(T)$ refers to a set of elements of type T. We use $Set(Numeric)$, $Set(Nominal)$, and $Set(String)$ to type certain fields in EHR. Similarly, the type $Sequence(T)$ defines a sequence whose elements are of type T. A record type is defined as $RT = T_1 \times T_2 \cdots \times T_n$. It defines the structure of an ordered collection of n items, where the type of i^{th} item is T_i. Each T_i may be either a simple or compound type. For practical purposes, we associate each field of a record with a typed attribute name (sometimes called "field name"). Hence, the definition of each record type declares the number of fields it has, their names, and their types. If r is a record of type RT, then $r. < Attribute_i >$ refers to the value associated with the i^{th} attribute of type T_i. We may view a record as a "multi-variate" vector. Each one these higher order types inherits the operations from its corresponding mathematical definition. Thus, set equality, and set-theoretic operations are inherited by the set type, while operations on type T are also

Table 2.4: Patient Records in a Hospital

Patient ID	**Name**	**Age**	**Gender**	**Date_Admi**
P1M1	Peter North	32	Male	15/May/1978
P2F1	Caroline Sue	43	Female	22/October/2010
P3F2	An Tao	53	Female	30/December/1998

available for the elements of type $Set(T)$. Since, equality and other operations are defined on the elements of T_i, we can carry these operations to record level. Two records r and s of type RT can be compared "component-wise". Hence, $r = s$ if and only if $r_i = s_i$. Table 2.2 shows five records of type $Categorical \times Categorical \times Range \times Range \times Interval$. Because of the recursiveness in type definition, we can introduce record types for the commonly arising attributes, such as *Date*, *Address*, in formalizing the record types that arise in health care applications. In Table 2.4 three patient records are shown. In this table, the fifth column attribute *Date_Adm* has three fields $Day/Month/Year$. The type of Day is *Range*, and its domain is $[1, 31]$. The type of $Month$ is "Ranked Categorical" with domain "the calendar months listed in the calendar order". The type of $Year$ is *Range*, whose domain is determined by the application domain. As an example, the domain $[1900, 2999]$ might serve the purpose of hospital records for all patients. In a similar way, we can define *Address* attribute as a record type. Records of different types are not comparable, however fields of same type of such records can be compared.

For the analyses that we focus in thesis, the attribute types that we have defined seem sufficient to model EHRs. They can be mixed in many ways to support robust models. Some examples of compounding the types are the following:

- The type $Set(Nominal)$ can model sets of drugs prescribed to a patient in the EHR.
- The type $Sequence(Record)$ can be used to model clinical sequences of a patient.
- The type $Set(RT)$ may be used to model the set of records of a specific record type RT.

In Chapter 4 we will discuss methods to compare and compute measures of similarity between sets of types real and sets of type nominal (supported by Ontology), because we use them in our analyses.

Chapter 3

Modeling Electronic Health Records

In this Chapter we first identify the category of users of EHR database. The EHR structure that we propose will be based on their needs, and record data needed by them using the types discussed in Chapter 2. Next, we review three clinical datasets to identify the most commonly used attributes. We ensure that our model includes them. Finally, we give the vector model of EHR that is structured into blocks where within each attribute within block include one category of information. Our model includes attributes necessary to model environmental and social aspects of a patient. Such attributes play a crucial role in research related to health surveillance (Birkhead et al., 2015), general population health (Kruse et al., 2018), and infectious disease control (Babcock, Beverley, Cowell, & Smith, 2021).

3.1 Attribute Variety for Serving Different Categories of Users

In this section we motivate why a large number of attributes with different types are necessary, from the perspective of user categories and their requirements.

- *Researchers*: A researcher (or a research group) needs a large amount of data that is relevant to the specific research goals. There are many research areas in healthcare domain, some of which are regarded more seasonal than others. Some of these are the following:

(1) Study of drug-drug similarity for drugs administered to patients in different diseases, such as cancer, diabetics, hypertension, infectious diseases, and mental disorders is an active research area. The similarity itself may be based on chemical structure in drugs, drug-drug interactions, side effect on patients, gene ontology, and ATC codes (Ferdousi et al., 2017; L. Huang, Luo, Yang, Wu, & Wang, 2021).

(2) Study on patient-patient similarity can be done in many ways. Some of these are similarity assessment with respect to disease type, set of drugs taken, allergy types, clinical visit profiles, and social aspects such as living and cultural issues. Each method requires a specific set of attributes.

(3) Study of infectious diseases or pandemic data with regard to their origin, the severity of its effect on geographical regions, and type of humans most affected within regions require large amount of environmental and geographical information in addition to disease-related information. So, attributes that model environmental and contextual aspects are necessary to be included in the EHR.

In summary, a variety of attributes are required to conduct EHR-based research.

- *Patients*: In general, patients want some privacy for their data, want to have some control over which data to share with whom, and above all want to have ready access and easy to understand data formats. Although some of these requirements belong to data display and data control, it is essential that these attribute types are chosen to facilitate such requirements. The patients should get overall assessments of their health trends based on their anonymous demographic information. When in different seasons or living in different areas, they should also be notified the seasonal influenza or ongoing spread of certain kind of virus. Depending on their age, gender and ethnic group, they should also be aware of their susceptible diseases and measures to defend themselves. The system should be responsive when the patients are in certain kind of needs such as emotional problems or lack of immunity, the system should be able to help the patient to do fundamental mental health checks and help them to set health goals or recommend them to turn to a physician. They should be able to review their clinical records.

- *Physicians*: They must have access to the EHRs of all patients under their care, although access rights may be contextually restricted to subsets of the actions *read, write, copy.* Clinical records, medication lists, and medical/diagnosis/treatment history of patients under the care of a physician should be made available to the physician. Hence, most of EHR information, other than those not allowed by a patient, will be available to the physician.

- *Nurses*: Attributes that pertain to patient's clinical data, contact information (physicians, emergency staff, and patient authorized personnel) for emergency situations, and patients's scheduled visits are essential for nurses.

- *Emergency Personnel*: This group of people should be given access to patient personal safety requirements (allergy, drugs) and essential cultural aspects in order to provide respectful care to the patient. The code of conduct as prescribed by the healthcare providers and the wishes of the patient should be followed without fail. So, attributes that are necessary to store this information in the EHR must be carefully screened and selected.

- *Administrators*: This group includes personnel in patient admission office, healthcare insurance providers, and system administrators who main the EHR system. System administrators need to have the power to monitor the overall security/privacy levels of the system, also keep the system users well-informed of new system features. They must interact with other healthcare users of the system to update overall trends and ongoing diseases in a certain area, and broadcast them to all eligible users in alerting some emergency healthcare measures that might be instituted.

The above list is neither exhaustive nor complete in every detail. Consequently, the choice of attributes, regardless how many and how diverse, may not sufficiently model an EHR. Consequently, a periodic review of EHR structure seems necessary.

3.2 Clinical Records Review

In this section we review the structure of three clinical records maintained in different hospitals, and bring out the attributes, their types, and their effectiveness for healthcare analysis. We observed that the “units” of measurement, say for vital signs such as glucose level, blood pressure, and pulse, are mentioned only as part of “metadata”, and not part of the records. In this thesis we follow the same convention.

3.2.1 Cleveland Clinic Heart Disease Dataset

Coronary Heart Disease (CHD) is the most common form of cardiovascular disease, with approximately 30% of patients dying after their first CHD event (Janosi et al., 2017). It can be very dangerous. Currently, the diagnosis of CHD is mainly based on invasive coronary angiography diagnosis method, which can be costly and may harm the patients as well. Although there are other less invasive diagnostics methods, the accuracy of those methods only ranges between 35%-75%. Thus it is possible to develop a computer-aided diagnostic method that can combine results of these non-invasive tests with other patient attributes to raise the diagnostic accuracy and eventually replace the invasive diagnostic methods.

The Cleveland clinic heart disease dataset is fetched from the UCI official site (Janosi et al., 2017). It contains 76 attributes and 14 of them are used for analysis, as shown in Table 3.1. Apart from age and gender, it also includes chest pain type (cp), which indicates the underlying heart problem, and the blood pressure levels are also often associated with the risk and diagnosis of CHD. For people in their 60s, a 10 mm Hg lower in systolic blood pressure caused about 20% lower risk of CHD. High cholesterol level can cause arteries become narrowed, thus reduce blood flow to the heart and increase the risk of getting heart diseases. If fasting blood sugar level is $<$ 70 mg/dL or $>$ 100 mg/dL the risk of getting CHD is increased. The report says that for analysis purposes the most *effective* set of attributes, among those shown in Table 3.1, are “age”, exercise-induced angina status (“exang”), ST/heart rate slope categories (“slope”), and stress scintigraphy results (“thal”).

3.2.2 Heart Failure Clinical Records Dataset

Cardiovascular diseases are very dangerous, which kill about 17 million around the globe each year (Mc Namara, 2019). Adults aged 65 and older are more likely than younger people to suffer from cardiovascular disease. The dataset in Table 3.2 is taken from (Chicco & Jurman, 2020). It contains 13 attributes, among which the “ejection_fraction” and “serum_creatinine” are found to be the most *effective* ones. It is reported that when

Table 3.1: Cleveland Heart Disease Dataset

Attribute	Measurement scale	Definition	Categories
age	Interval	Age in years	-
gender	Nominal	gender in nominal	(1) Male; (0) Female
cp	Nominal	Chest pain type	(1) Typical angina; (2) Atypical angina; (3) Nonanginal pain; (4) Asymptomatic
trestbps	Interval	Peak exercise systolic blood pressure (in mmHg on admission to the hospital)	-
chol	Interval	Serum cholesterol in mg/dL	-
fbs	Nominal	Fasting blood sugar $>$ 120 mg/dL	True/False
restecg	Nominal	Resting electrocardiographic results	(1) Normal; (2) Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of $>$ 0.05 mV); (3) probable or definite left ventricular hypertrophy by Estes' criteria
thalach	Interval	Maximum heart rate achieved (bpm)	-
exang	Nominal	Exercise included angina	Yes/No
oldpeak	Interval	ST depression included by exercise relative to rest	-
slope	Ordinal	The slope of the peak exercise ST segment	(1) Upsloping; (2) Flat; (3) Downsloping
ca	Interval	Number of major vessels (0-3) coloured by fluoroscopy (for calcification of vessels)	-
thal	Nominal	Results of nuclear stress test	(3) Normal; (6) Fixed defect; (7) Reversible defect
num	Nominal	Diagnosis of heart disease (angiographic disease status)	(1) Normal: <50% diameter narrowing; (2) CAD>50% diameter narrowing

Table 3.2: Heart Failure Clinical Records Dataset

Attribute	Measurement scale	Definition	Categories
age	Interval	Age in years	-
anaemia	Nominal	Decrease of red blood cells or hemoglobin	(1) Yes; (0)No
high__blood__pressure	Nominal	If the patient has hypertension	(1) Yes; (0)No
creatinine__phosphokinase	Interval	Level of the CPK enzyme in the blood (msg/L)	-
diabetes	Nominal	If the patient has diabetes	(1) Yes; (0)No
ejection__fraction	Interval	Percentage of blood leaving the heart at each contraction (%)	-
gender	Nominal	Gender in nominal	(1) Male; (0) Female
plateltes	Interval	Platelets in the blood (kilo platelets/mL)	-
secum__creatinine	Interval	Level of creatinine in the blood (mg/dL)	-
serum__sodium	Interval	Level of sodium in the blood (mEq/L)	-
smoking	Nominal	If the patient smokes	(1) Yes; (0)No
time	Interval	Follow-up period (4-285 days)	-
death__event	Nominal	If the patient died during the follow-up period	(1) Yes; (0)No

predicting the survival of patients with heart failure, using only these two attributes can even get better results than using all attributes together.

3.2.3 HCC Survival Dataset

Hepatocellular Carcinoma (HCC) represents more than 90% of primary liver cancers, which is the sixth most frequently diagnosed cancer in the world (McGlynn & London, 2011). This research concludes that liver cancer cannot be detected by blood tests alone. Depending only on the biological variability causing the disease, clinicians may not give each individual patient the proper treatment. So, different records from different data sources and with different data types are being collected for different types of analysis, hoping that the quality of treatment can be improved through investigation closer to the real-world

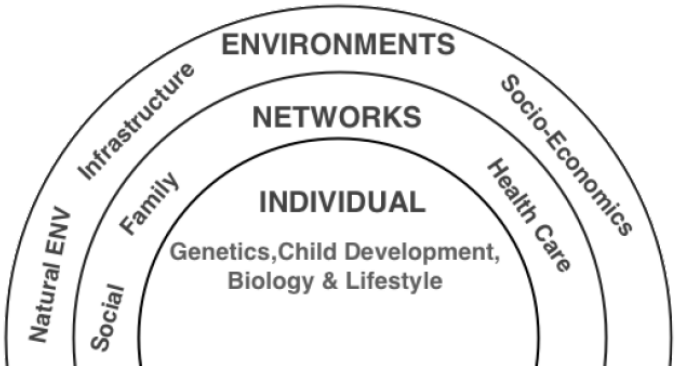

Figure 3.1: Patient-centric Health Care Model

situations. The datasets in (Miriam Seoane Santos, 2017) contain data collected from 165 patients diagnosed with HCC. Each record has 50 attributes, which are divided into the three groups *living habit and demographic attributes*, *diagnosis attributes* and *blood test results attributes.* The first 10 attributes in HCC survival dataset are about the patient's life habits and demographic information. Primarily these attributes are the patient's profile on living habits, family and personal preference-centric attributes. We expand and enrich this dataset in (McGlynn & London, 2011) with more attributes from "Patient-centric Healthcare Model" proposed in Section 3.2.4. The 16 diagnostic attributes are shown in Table 3.3. They are more specific to the diagnostic history of a patient. It provides information about symptoms, other chronic diseases, and more common diseases such as diabetes and hypertension which may have an impact on getting liver cancer.

Table 3.4 lists the 24 attributes for blood test results. They are the most common attributes in use for analysis. As HCC cannot be diagnosed by "routine" blood tests, these specific blood test attributes are needed for HCC analysis.

3.2.4 Patient-centric Healthcare Model: Attributes for Living Habit and Demographic Information

The patient-centric healthcare model proposed in (K. Wan & Alagar, 2015) is shown in Figure 3.1. In this layered ring model, the inner most layer includes the personal and heathcare attributes specific to an individual patient. The middle layer includes the attributes that model the family, social, and healthcare networks that are related to the health determinants of the patient. The outermost layer includes the attributes that describe the living

Table 3.3: Diagnostic Attributes In HCC Survival Dataset

Name	Measurement scale	Abbreviation	Categories	Missing Values(%)
Symptoms	Nominal	Symptoms	1=Yes; 0=No;	10.91
Cirrhosis	Nominal	Cirrhosis	1=Yes; 0=No;	0
Diabetes	Nominal	Diabetes	1=Yes; 0=No;	1.82
Hemochromatosis	Nominal	Hemochro	1=Yes; 0=No;	13.94
Arterial Hypertension	Nominal	AHT	1=Yes; 0=No;	1.82
Chronic Renal Insufficiency	Nominal	CRI	1=Yes; 0=No;	1.21
Human Immunodeficiency Virus	Nominal	HIV	1=Yes; 0=No;	8.48
Nonalcoholic Steatohepatitis	Nominal	NASH	1=Yes; 0=No;	13.33
Esophageal Varices	Nominal	Varices	1=Yes; 0=No;	31.52
Splenomegaly	Nominal	Spleno	1=Yes; 0=No;	9.09
Hypertension	Nominal	PHT	1=Yes; 0=No;	6.67
Vein Thrombosis	Nominal	PVT	1=Yes; 0=No;	1.82
Liver Metastasis	Nominal	Metastasis	1=Yes; 0=No;	2.42
Radiological Hallmark	Nominal	Hallmark	1=Yes; 0=No;	1.21
Encephalopathy degree*	Ordinal	Encephalopathy	1=None; 2=Grade I/II; 3=Grade III/IV;	0.61
Ascites degree*	Ordinal	AHT	1=None; 2=Mild; 3=Moderate to Severe;	1.21

Table 3.4: Test Results Attributes In HCC Survival Dataset

Test Name	Attribute	Measurement scale	Categories	Missing Values(%)
Hepatitis B Blood Test	Hepatitis B Surface Antigen	Nominal	1=Yes; 0=No;	10.3
	Hepatitis B e Antigen	Nominal	1=Yes; 0=No;	23.64
	Hepatitis B Core Antibody	Nominal	1=Yes; 0=No;	14.55
Hepatitis C Blood Test	Hepatitis C Virus Antibody	Nominal	1=Yes; 0=No;	5.45
Alpha-fetoprotein Blood (AFP) Test	Alpha-Fetoprotein (ng/mL)	Continuous	1.2-1810346	4.85
PT/INR Test	International Normalised Ratio*	Continuous	0.84-4.82	2.42
MRI	Number of Nodules	Integer	0-5	1.21
	Major dimension of nodule (cm)	Continuous	1.5-22	12.12
Complete Blood Count (CBC)	Haemoglobin (g/dL)	Continuous	5-18.7	1.82
	Leukocytes (G/L)	Continuous	2.2-13000	1.82
	Platelets (G/L)	Continuous	1.71-459000	1.82
MCV Level Blood Test	Mean Corpuscular Volume(fl)	Continuous	69.5-119.6	1.82
Albumin Blood Test	Albumin (mg/dL)	Continuous	1.9-4.9	3.64
Total Bilirubin (Blood) Test	Total Bilirubin (mg/dL)	Continuous	0.3-40.5	3.03
Alanine Aminotransferease (ALT) Test	Alanine transaminase (U/L)	Continuous	11-420	2.42
Aspartate Aminotransferase (AST) Test	Aspartate transaminase (U/L)	Continuous	17-553	1.82
Gamma-glutamyl Transferase (GGT) Test	Gamma glutamyl transferase (U/L)	Continuous	1=Yes; 0=No;	1.82
Alkaline Phosphatase (ALP) Test	Alkaline phosphatase (U/L)	Continuous	1=Yes; 0=No;	1.82
Total Protein Test	Total Proteins (g/dL)	Continuous	17-553	6.67
Creatinine tests	Creatinine (mg/dL)	Continuous	0.2-7.6	4.24
Bilirubin test	Direct Bilirubin (mg/dL)	Continuous	0.1-29.3	26.67
Iron Blood Tests	Iron (mcg/dL)	Continuous	0-244	47.88
Blood Oxygen Level Test	Oxygen Saturation (%)	Continuous	0-126	48.48
Ferritin Test	Ferritin (ng/mL)	Continuous	1=Yes; 0=No;	48.48

demographic situation of the patient. The patient determines the number of attributes and their types in each layer. By including such attributes in the EHR we achieve a level of patient-level completeness and acceptance, which will lead to patient-level acceptance of analysis results. Given that "living habit and demographic attributes" are necessary for HCC analysis, we can enrich the set of attributes necessary for an effective analysis by including the attributes of the elements in the patient-centric healthcare model in (K. Wan & Alagar, 2015). This enriched set of attributes is shown in Table 3.5. Based upon similarity measures on attributes that model living habit, family network, and infrastructure support we agree with the opinion (Babcock et al., 2021; Birkhead et al., 2015; Kruse et al., 2018).that it is possible to conduct a more comprehensive research on patient-patient similarity assessment, and infer how they collectively influence the spread of certain disease types in the general population.

3.3 Proposed EHR Model

From the above study we conclude that in existing hospital datasets, the set of attributes is not necessary a fixed set. The set of attributes seem to vary and grow as and when the analysis team faces new challenges. So, we need a rich EHR model that is expandable. After reviewing the different needs for different groups of users involved in medical system, and learning about different real-world datasets that are being used by researchers in different analyses, we model EHR as a vector of attributes as shown in Figure 3.2. The EHR vector has 5 parts, which respectively model the personal information, disease information, drug information, clinical visits information, and social aspects information of a patient. The set of attributes for each part may be selected to fit the overall goal of EHR use in an organization. Each patient has a unique EHR, which can be identified by a combination of patient identifier (hospital card, medicare card, insurance card). Within each field, the attributes are ordered (in a certain way), and each attribute has a type. The internal representation of EHR may vary from one implementation to another, however for us the logical structure Figure 3.2 is referred through its unique identifier. An authenticated user will be given access to a *view* of EHR as determined by the access control authentication policy implemented in the organization. The projected EHR view may be partial or total,

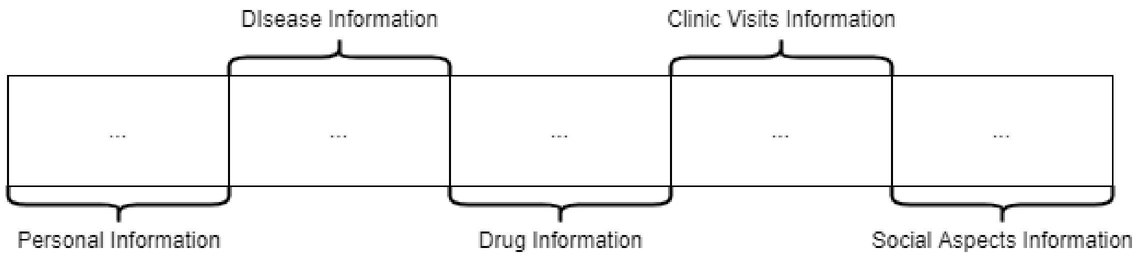

Figure 3.2: EHR Vector Structure

and has a unique *pseudo ID* through which the updated view may be carried over the system EHR. A partial view is a *projection* of the full EHR on the attributes that the user is allowed to view/copy/modify according to access control validation results. The projected order of the parts and the order of attributes within a part are consistent with the orders in the EHR. That is, the set of attributes in the projected EHR is a *sub-sequence* of the EHR sequence structure. We also assume that for a group of users, say research analysis group, the projected view of EHRs is according to the group-level access control validation results. So, in this thesis we focus only on similarity analysis of projected EHRs, assuming that safety and privacy policies are enforced by the system, and through the pseudo-Id the results of analysis can be carried over to the system EHR.

3.3.1 Personal Information

Table 3.6 lists the most commonly used attributes to model the personal information. Attribute-level praivacy/security, as agreed with a patient may be enforced so that personal information is shared only with authenticated individuals, organizations and groups. Patient ID, if it serves as the key to EHR, and other attributes not authorized by the patient must be replaced with other anonymous identifier to avoid the potential leak on the patient's personal information.

3.3.2 Disease Information

A patient may have one more diseases, each at different stages of affliction. So, attribute names are disease names and their specific attributes. Table 3.7 lists a sample set of diseases for a patient. It includes disease names, attribute type for disease, and disease type.

3.3.3 Drug Information

Table 3.8 lists some of the important attributes for drugs included in the patient EHR. The name of the drug, list of its generic names, the ATC codes for drugs, drug dosage and frequency of daily use are the essential attributes. A patient's EHR might be structured to relate the drug information for each disease type of the patient. A Drug Product Database (DPD) and drug ontology might be part of the system, which can be accessed by medical professionals and research analysts.

3.3.4 Clinical Visits Information

Some researchers (Z. Huang et al., 2013; van de Klundert, Gorissen, & Zeemering, 2010) have studied patient similarity based on the measure on clinical pathway adherence and similarity measure between patient traces. A patient trace is a non-empty sequence of clinical events $< e_1, e_2, \cdots, c_m >$ performed by a particular patient. An event e_i is a pair $(Name, Timesteamp)$, where $Name$ denotes the "clinic name visited" and the $Timestamp$ is an interval $[Time\_in, Time\_out]$. By associating the clinical charts with such traces, they try to measure how similar patients are in getting their treatment and how they respond to treatments (medications) prescribed at each visit. The clinical information in an EHR, as sampled in Table 3.9, is to comprehensively aid such analysis. The sample information in the table contains the clinic visited (facility), data and times of coming into clinic and leaving, drugs and dosages administered at each visit, the name of attending physician (may be different from the ones in the personal profile of the patient), and specifics of diagnosis.

3.3.5 Environmental and Social Aspects Information

Attributes for social aspects governing a patient listed in Table 3.10 are related to the outer layers in Figure 3.1. Including these attributes in the EHR will enhance patient-centric medical care. These attributes may change as and when the patient's social networking or/and environmental situations change.

Table 3.5: Living Habit And Demographic Attributes in HCC Survival Dataset

Name	Attribute Type	Categories/Values
Gender	Nominal	1=Male; 0=Female
Alcohol	Binary	1=Yes; 0=No;
Smoking	Binary	1=Yes; 0=No;
Obesity	Interval	1=Yes; 0=No;
Age at diagnosis	Enumerated	$20, \cdots, 95$
Grams of Alcohol per day	range	$[0, 500]$
Income Level	Nominal	$1, \cdots, 6 \vert\vert 1 :\leq 20k; 2 : (20k - 40k]$; $3 : (40k - 60k]$; 4: $(60k - 80k]$; $5 : (80k - 100k]; 6 :> 100k$
Literacy	Nominal	$1, \cdots, 4$!! 1 : High School; 2 : Vocational College 3 : Undergraduate4 : Post Graduate
Harmone Type	Nominal	$1, \cdots, 6$!!1 : $T_3(Thyroid); 2 : T_4(Thyroid)$ 3 : Melatonin; 4 : testosterone; 5 : Estrogen : 6 : Cortisol
Marital Status	Nominal	$1, \cdots, 4$!! 1 : Single; 2 : Living Together 3 : Married4 : Divorced
Family Structure	Nominal	$1, \cdots, 4$!! 1 : Single Parent; 2 : Extended 3 : Step Family4 : Conventional
Living Neighbourhood	Nominal	$1, \cdots, 6$!! 1 : Near Hospital; 2 : Poor Public Transportation; 3 : Near Industry Park; 4 : Near Social Clubs; 5 : No Public School : 6 : Near Fire Station
Treatment Preference	Nominal	$1, \cdots, 4$; !! 1 : No Medical Devices; 2 : Vegetarian Nutrition; 3 : No Blood Transfusion 4 : Avoid Blood Products

Table 3.6: Personal Information Attributes

Attribute Name	Type	Sample Attribute Values
Patient ID	String	- ALAV256798
First Name	Nominal	Smith
Last Name	Nominal	William
Gender	Nominal	Male
Age	Integer	27
Marital Status	Nominal	Single, Married, Divorced
Professional Status	Nominal	Employed, Out of Work, Retired
Residential Address	Record	[1026|2762 Treeline|Montreal|Sudbury|K5L 3H6| Canada]
Hospital Card	String	MUHC37965
Nationality	Nominal	Canada
Ethnic Group	Nominal	Native Indian, Asian
Education Level	Nominal	Elementary School, High School, College, University
Allergic Type	Nominal	Pet Allergy, Pollen Allergy, Food Allergy
Blood Type	Nominal	AB
Family Doctor	Nominal	Joseph Hartman
Health Situation	Nominal	Anxiety, Depressin, Hypertension
Primary Physician	Nominal	Mark Zabo
Income Level	Nominal	Low, Medium, High, Very High

Table 3.7: Disease Information Attributes

Disease Name	Attribute Type	Disease Type	Disease Code
Diabetes	Nominal	Type 2 hyperglycemia	ICD-10-CM Diagnosis Code E11.65
Anaemia	Nominal	Iron Deficiency	ICD-10-CM Diagnosis Code D50.9
Blood Pressure	Nominal	Low	ICD-10-CM Diagnosis Code R03.1

Table 3.8: Drug Information Attributes

Attribute Name	Attribute Type	Sample Attribute Values
Drug Name	Nominal	Gemzar Infugem
Generic Name	Nominal	Gemcitabine
Protein Chemical Formula	Nominal	C9H11F2N3O4
Synonyms	Nominal	Gemcitabin
ATC Codes	Nominal	L01BC05 — Gemcitabine
Frequency/day	Interval	[2,4]
Dosage	Nominal	300mg
Side Effect	Nominal	Headache, Vomiting, DiarrheV
Remarks	String	Can harm the fetus if taken by a pregnant woman

Table 3.9: Clinical Traces Attributes

Attribute Name	Type	Sample Attribute Values
Patient ID	String	YNLI46432
Visit Date	Record	[04|25|2022]
Visit Times	Interval	$[11:30, 14:15]$
Facility	Record	[Eye Clinic, Verdun Hospital]
Test Results	Record	[Glucose:7.2|BP:140/90|Normal Vision]
Diagnosis	Nominal	Mild Cataract
Physician	String	Brian Stewart
Severe Level	Nominal	Low
Drugs (ATC Codes)	Nominal	L01BC15
Dosage	Nominal	Normal
Strength	Nominal	1g /25mL

Table 3.10: Environmental and Social Aspects Attributes

Attribute Name	Type	Sample Attribute Values	Remarks
Water Quality	Enumerated	Poor, Pure	Canal Water
Air Quality	Enumerated	Allergic	Wilderness
Health Center	Nominal	Metro CLSC	Not Close
Guardian	String	Mary Khan	Phone:416-354-7689
Networking	Nominal	City Center	Senior Group
Neighbourhood	Nominal	Lower Income Group	Near Industrial Park
Law & Order	Nominal	Safe	Good Policing

Chapter 4

Similarity Functions: A Critical Review and Proposal of New Functions

In this chapter we critically review the similarity functions used by researchers in the fields of healthcare, psychology, and in biomedical informatics, and after comparing their merits we choose two of them to modify and extend for our similarity study. We propose one new similarity function that combines semantic distances in Ontology and two new similarity functions for sets of concepts supported by Ontology. After introducing the basic notion of "similarity" and object representation for similarity study, we discuss a set of criteria for comparing/evaluating similarity functions. After that, we systematically review similarity functions that are very relevant to our thesis goal, and propose our methods. Several examples and case studies are included to bring out the performance and merits of our methods. Finally, we discuss user-level semantics which can be integrated in the similarity functions proposed by us.

4.1 Similarity and Object Modeling

The concept "similarity" has been studied for a long time (Tversky, 1977). It has played a fundamental role in *classification* and *clustering* of objects of importance in the study of economic behavior (von Neumann & Morgenstern, 1947), psychology (Eisler & Ekman, 1959; Reed, 1972), knowledge discovery from data (Tversky & Krantz, 1970), pattern recognition (Reed, 1972), information retrieval (Metzler, Dumais, & Meek, 2007; Tombros & Rijsbergen, 2004), and meteorological studies (Mo, Ye, & Whitefield, 2013). Recently (Chan et al., 2010; Sun, Wang, & Edabollahi, 2012; Zhang et al., 2014) similarity measures have been used on Electronic Health Records (EHR to investigate patient similarity, drug-drug similarity, and mortality rate prediction (J. Lee et al., 2015). In spite of the sound mathematical foundations for constructing a variety of similarity functions, and the long history of their applications to a variety of fields, determining the best similarity function for any specific application remains only as an experimental issue. A function that is experimentally found to perform well for one application may fail to produce good results for another application. The major reason for this deficiency can be traced to (1) the lack of appropriateness of the set of *features* (*attributes*) selected to model an object, and (2) the selection of similarity function without prescribing the criteria to be met in accepting the measured similarity. Similarity functions constructed by injecting domain-specific semantics (Alsaig, 2013; El-Sappagh et al., 2018; Harispe et al., 2014) have been found to have the potential to perform well in similarity-based clustering, ranking, and classifications. In addition, prescribing the necessary criteria to be met for *fairness* in computing measures (Alsaig, Alagar, Mohammad, & Alhalabi, 2017) leads to a theoretical, rather than just experimental, basis for comparing similarity measures and accepting the "best" one. In this chapter, we first review fundamental definitions and give a list of most popular measures of similarity. Next, we bring out the essential aspects of the comparative study from literature in order to motivate the types of similarity functions that have been used in health care analysis. Next, we motivate the properties of similarity functions that are likely to benefit our analysis goals. Finally, we explain the construction steps of a semantic-based similarity function that we will use in this thesis.

4.1.1 Object Representation for Similarity Study

The two approaches to study similarity relations of objects are based on *geometric modeling* and *set-theoretic modeling* of objects (Tversky, 1977).

In geometric modeling, an object is conceptualized as a point in a higher-dimensional space, and similarity (dissimilarity) between objects is studied in terms of distance metrics. A distance function ρ assigns to every pair of points (objects) x and y a non-negative number which satisfies the following three axioms:

- *Minimality:* for $x \neq y$, $\rho(x, y) \geq \rho(x, x) = 0$,

- *Symmetry:* $\rho(x, y) = \rho(y, x)$, and

- *The Triangle Inequality:* $\rho(x, y) + \rho(y, z) \geq \rho(x, z)$.

If the dimensionality of the modeling space is n, the underlying assumption is that every object x is represented as a point with n coordinates $(x_1, x_2, \cdots, x_n)$ in the Cartesian co-ordinate system, where x_is are real numbers. The well-understood Euclidean distance formula,

$$\rho(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$$

satisfies the above three axioms.

In set theoretic modeling, an object is conceptualized as a set of *features* (*attributes*), and similarity of objects are studied in terms of functions defined on union, intersection, and difference of feature sets. Let A, B and C respectively denote the set of features of objects a, b, and c. Then, Tversky (Tversky, 1977) defines the "generic similarity" function

$$S_{TV}(a, b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha f(A \setminus B) + \beta f(B \setminus A)}, \quad \alpha, \beta > 0, \tag{1}$$

where$f(X) > 0, X \neq \emptyset,$ and$f(\emptyset) = 0$. In this seminal work Tversky (Tversky, 1977) proves the following results:

- For non-directional similarity assessment, it is required to assess the degree to which "objects a and b are similar to each other". In this form, $S_{TV}(a, b) = S_{TV}(b, a)$ holds if $\alpha = \beta$.

- For directional similarity assessment it is required to assess the degree to which "objects a is similar to object b". In this form, $\alpha \neq \beta$, and $S_{TV}(a, b) \neq S_{TV}(b, a)$.

Thus, the function $S_{TV}(a, b)$ defined in Equation 1 is very general. For both symmetric and asymmetric forms of similarity, we can verify that the similarity measure is in the interval

$[0, 1]$. With the choice $f(X) =\mid X \mid$, we can rewrite Equation 1 as

$$S_{TVM}(a, b) = \frac{|(A \cap B)|}{|(A \cap B)| + \alpha|(A \setminus B)| + \beta|(B \setminus A)|}, \quad \alpha, \beta > 0 \tag{2}$$

The measure $f(X) =\mid X \mid$, that calculates the number of elements in the set X, may be replaced by "weighted sum" $\Sigma_{i=1,n} w_i$. where $\mid X \mid= n$, and $w_1, w_2, \cdots, w_n$ are the weights assigned to the elements of set X according to a predefined priority scheme.

Example 3. *Let $A = (0, 1, 3, 5, 8)$ and $B = (0, 2, 4, 5, 8, 9)$ denote the cartesian coordinate representations of two objects a and b in their geometric modeling. Because their dimensions are different, we cannot apply the ρ function. However, we can apply S function when the set of coordinates model each object in the "set-theoretic" modeling. We have $A \cap B = \{0, 5, 8\}$, $A \setminus B = \{1, 3\}$, $B \setminus A = \{2, 4, 9\}$. Substituting the cardinalities of these sets in the formula in Equation 2, we get*

$$S_{TVM}(a, b) = \frac{3}{3 + 2\alpha + 3\beta}; \qquad S_{TVM}(b, a) = \frac{3}{3 + 3\alpha + 2\beta}$$

For $\alpha = 2$ and $\beta = 1$, we have

$$S_{TVM}(a, b) = \frac{3}{10}; \qquad S_{TVM}(b, a) = \frac{3}{11}; \qquad S_{TVM}(a, b) \neq S_{TVM}(b, a)$$

For $\alpha = \beta = 1$ we have

$$S_{TVM}(a, b) = \frac{3}{8}; \qquad S_{TVM}(b, a) = \frac{3}{8}; \qquad S_{TVM}(a, b) = S_{TVM}(b, a)$$

Remark 1. *From the above basic results, we can draw the following conclusions.*

1. The distance function $\rho(a, b)$ has the following properties.

- *It is small if objects a and b are "close", and it is large if the objects are "far apart".*
- *It is 0 if the objects are the same,*
- *In general (unless normalized), it has value in $[0, \infty]$.*

2. The similarity function $S_{TVM}(a, b)$ has the following properties.

- *It is large if objects a and b are "close", and it is small if the objects are "far apart".*

- *It is* 1 *if the objects are the same,*

- *It has value in* $[0, 1]$.

3. The function $\rho(a, b) = 1 - S_{TVM}(a, b)$ *is non-negative and is a metric (Chierichetti & Kumar, 2015). This result enables us to move from symmetric set similarity function to distance function, and vice versa.*

Influenced by the generic nature of the set similarity function, many generalized distance metrics for similarity assessment have also been defined. These are discussed in the following section. In our analysis we will be dealing with EHRs of heterogeneous types, and depending upon the analysis we will be using both dimensional and set-theoretic similarity calculations.

4.2 Similarity Functions and Their Properties

In this section we classify the most commonly used functions to compute similarity into four groups. For evaluating them, we informally argue from a mathematical perspective, and follow the criteria that have been experimentally validated (Alsaig, 2013). We use vector notation, both for the geometric modeling and set theoretic modeling of objects. So, an object a is a *record* type vector $A = [a_1, a_2, \cdots, a_n]$, $n \geq 1$. We also assume that all objects under consideration has n features. The vector notation for feature sets is only to simplify the notation in accessing elements, and the ordering in the vector is arbitrary although once chosen the same order is to be respected.

4.2.1 Criteria of Evaluation

A similarity function used in analyzing health records must be easy to compute, must produce normalized result in a bounded interval, say $[0, 1]$, must be sensitive to discriminate between small values in the results, and must be unbiased. These four requirements are defined and explained below.

Simplicity: A similarity function should be simple to calculate and easy to apply, in order to minimize computational complexity. So, the function definition must avoid computing "square roots, and powers". Ideally, it should be a linear function involving a minimum number of operators, excluding division but for normalization. With simplicity, it is easier to integrate it with semantic constraints and achieve timeliness and efficiency in similarity calculation.

Normalized Output: Normalization means that the similarity measures are adjusted to a common scale. Consequently, the results are bounded. If the similarity function does not produce normalized outputs, the result may include unexpectedly sparse numbers which cannot be understood and fairly ranked. Most similarity functions calculate the similarity between two vectors by aggregating the sub-scores of each component pairs of vectors. One example of aggregation is the "additive aggregation", in which all sub-scores are added to compute the similarity measure of vectors. Having differences in the range of values of attributes will affect the fairness of the results, in the sense that scores of attributes that have high values will influence the total measure. By normalization of each attribute value, this unfairness can be eliminated.

Sensitivity to Discriminate: This requirement is applicable for real valued attributes. A similarity function must be able to detect small differences between health records in order to classify them correctly. As an example, consider drug-drug similarity analysis. Let $A = [a_1, a_2, \cdots, a_n]$ and $B = [b_1, b_2, \cdots, b_n]$ be two drug records, where attributes a_i and b_i denote the presence of a critical chemical in the records. Let us assume $a_i = 0.51$, and $b_i = 0.513$ be the dosage of the chemical in the records. The chosen similarity function is sensitive to small difference if it produces different scores for a_i and b_i. To ensure fair classification it is necessary that the function is able to discriminate such small differences to compute different scores.

Unbiased: This requirement is applicable for real valued attributes and those for which a distance metric can be defined. Let $A = [a_1, a_2, \cdots, a_n]$, $B = [b_1, b_2, \cdots, b_n]$, and $C = [c_1, c_2, \cdots, c_n]$ be three vectors, and we want to calculate the similarity between (A, B) and between (A, C). Assume $|c_i - b_i| = |a_i - c_i|$, where $|X - Y|$ is the absolute value function. For example, $b_i = c_i + h$, and $a_i = c_i - h$. We say the similarity function is unbiased, if it assigns the same score to the pairs (c_i, b_i) and (a_i, c_i). However, this requirement may be overridden by some semantic constraints, such as "lower value is better" (for diabetic attribute) or “higher is better” (for some chemical compounds).

Table 4.1: Group 1: Distance Metric Measures - Values in the range $[0, \infty)$

Name/Formula	Name/Formula
(A) Euclidean Distance: L_2: $d_{Euc} = \sqrt[2]{\sum_{i=1}^{n} \lvert a_i - b_i \rvert^2}$	(B) City Block: L_1: $d_{CB} = \sum_{i=1}^{n} \lvert a_i - b_i \rvert$
(C) Chebyshev: L_∞: $d_{Cheb} = max_{i=1,n}\{\lvert a_i - b_i \rvert\}$	(D) Minkowski: L_p: $d_{Mk} = \sqrt[p]{\sum_{i=1}^{n} \lvert a_i - b_i \rvert^p}$ p is a real number, $p \geq 1$
(E) Kulczynski: $d_{kul} = \frac{\sum_{i=1}^{n} \lvert a_i - b_i \rvert}{\sum_{i=1}^{n} min(a_i, b_i)}$	(F) Gower: $d_{gow} = \frac{1}{n} \sum_{i=1}^{n} \lvert a_i - b_i \rvert$
(G) Lorentzian: $d_{lor} = \sum_{i=1}^{n} \ln(1 + \lvert a_i - b_i \rvert)$	(H) Inner Product: $d_{IP} = \sum_{i=1}^{n} a_i b_i$

4.2.2 Distance Functions Used in Literature: Classification and Comparison

The distance functions for geometric modeling of objects are divided into two groups. The first group, listed in Table 4.1, includes all the distance functions that produce results in the range $[0, \infty]$. The second group, listed in Table 4.2, includes the distance functions that produce results in a bounded range. In all the formulas, A and B are vectors of the same length, and a_i and b_i are their coordinates in the i^{th} dimension in an n-dimensional space. These formulas have been used for a long time in a variety of applications (Mo et al., 2013; Reed, 1972; Tombros & Rijsbergen, 2004), where there is some pre-knowledge about feature values. The most commonly used distance functions in patient-similarity analysis are the cosine function (K) and a normalized version of inner product function (H) (J. Lee et al., 2015).

Based on the experimental evidence (Alsaig, 2013) and through inspection of the formulas of distance functions, the behavior of the similarity functions in Table 4.1 and Table 4.2 with respect to our four similarity measure criteria are summarized below.

- *Simplicity:* Functions (B), (C) and (H) are simple. All others require square root or logarithm, or division operations. However, none of these functions satisfy the other

Table 4.2: Group 2: Distance Metric Measures - Values in a bounded range $[0, K]$

Name/Formula	Name/Formula
(I) Sørensen: $d_{sor} = \frac{\sum_{i=1}^{n} \lvert a_i - b_i \rvert}{\sum_{i=1}^{n}(a_i + b_i)}$	(J) Soergel: $d_{sg} = \frac{\sum_{i=1}^{n} \lvert a_i - b_i \rvert}{\sum_{i=1}^{n} max(a_i, b_i)}$
(K) Cosine: $S_{cos} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}$	(L) Canberra: $d_{can} = \sum_{i=1}^{n} \frac{\lvert a_i - b_i \rvert}{a_i + b_i}$
(M) Jaccard: $d_{jac} = \frac{\sum_{i=1}^{n}(a_i - b_i)^2}{\sum_{i=1}^{n} a_i^2 + \sum_{i=1}^{n} b_i^2 - \sum_{i=1}^{n} a_i b_i}$	(N) Harmonic Mean: $d_{HM} = 2\sum_{i=1}^{n} \frac{a_i b_i}{a_i + b_i}$
(O) Dice: $d_{dice} = \frac{\sum_{i=1}^{n}(a_i - b_i)^2}{\sum_{i=1}^{n} a_i^2 \sum_{i=1}^{n} b_i^2}$	(P) Relative Change: $S_{RC} = \sum_{i=1}^{n} \frac{\lvert a_i - b_i \rvert}{max(a_i, b_i)}$

criteria.

- *Normalized Output:* All Functions in Table 4.2 satisfy this criteria.

- *Sensitivity:* Whenever the denominator in the function formula increases lot faster the numerator, normalization happens, however the ability to discriminate between small values decreases fast. The similarity functions (K), (M), and (O) are the only ones that satisfy this criterion.

- *Unbiased:* Just by choosing one pair attributes a_i, b_i that are symmetric with respect to c_i, and substituting $b_i = c_i + h$, and $a_i = c_i - h$ in the formulas we can measure the similarities between (a_i, c_i) and $b_i, c_i)$. It can be verified that only functions (A), (F), and (G) are unbiased.

In summary, all functions (A) to (P) implicitly assume that the attributes have numerical values and vectors are of the same length. In many healthcare applications, neither requirement can be fulfilled. Because, not all patient attributes will be numerical, and it is more common for patient data to be incomplete. Moreover, none of the distance functions meets all the four similarity function criteria. However, functions (A), (F), (I), (J), (L), and (P) meet three out of four criteria. But, the first two do not meet the normalization requirement. So, they are not suitable for our study. All the functions (I), (J), (L) and (P) lack only "unbiased" property. In both (I) and (J) the similarity of pairs of attributes are aggregated in one step. If we adapt these functions then we can not associate priority weights for attributes and calculate weighted aggregation, In functions (L) and (P) we can introduce scalars for assessing the similarity for each pair. The denominator of function (L) grows much faster than the denominator of function (P), which increases the precision required to discriminate between small values. So, function (P) seems the best one to choose and then specialize it to suit the semantic needs.

4.2.3 Set-theoretic Similarity Functions Used in Literature

The function in Equation 2 is the most general form of set-theoretic similarity function. It has been used in semantic web applications (Likavec, Lombardi, & Cena, 2015), and for molecular comparison in drug-drug similarity classification (Kunimoto1 et al., 2016), and in facial recognition problems (Reed, 1972). Some of the other set-theoretic similarity functions that are shown in Table 4.3 have been used in drug-drug similarity classification (Y. Huang et al., 2019; Zhang et al., 2014) studies. Many variations of these functions have been introduced to study ontology-based semantic similarity functions in biomedical applications (Girardi et al., 2016; Harispe et al., 2014; Mabotuwana et al., 2013).

Table 4.3: Set Theoretic Similarity Functions - In Addition to Formulas in Equation 1 and Equation 2

Name/Formula	Name/Formula
(JS) Jaccard: $JS(A, B) = \frac{\|A \cap B\|}{\|A \cup B\|}$	(HA) Hamming: $Ham(A, B) = 1 - \frac{\|A \Delta B\|}{\|[SS]\|}$
(AN) Andberg: $Andb(A, B) = \frac{\|A \cap B\|}{\|A \cup B\| + \|A \Delta B\|}$	(SD) Sorensen-Dice: $Dice(A, B) = \frac{2\|A \cap B\|}{\|A\| + \|B\|}$

In Table 4.3 the set SS denotes the "superset" that contains all attributes. That is, every set A and B that are compared is a subset SS. The *symmetric difference* Δ between two sets A and B is defined as

$$A\Delta B = (A \cup B) \setminus (A \cap B)$$

Remark 2. *Some of the properties of the set-theoretic measures include the following.*

(1) The attributes (set elements) can be heterogeneous.

(2) Not all records (vectors) need to have the same number of attributes or to have the same attributes.

(3) All similarity functions are simple to compute and have bounded values. Function $S(a, b)$ in Equation 2 is symmetric if $\alpha = \beta$. In this case, since $0 \leq S(a, b) \leq 1$, $1 - S(a, b)$ is a distance metric. Notice that $A\Delta B = B\Delta A$. Hence, all functions in Table 4.3 are symmetric. Only functions JA, HA, and AN have values in the interval $[0, 1]$. Thus $1 - S(A, B)$, where S stands for any one of these, is a metric. Hence, we can use JA, HA, and AN wherever distance metric is needed.

(4) All these functions are immune to "unbiased" property and have sensitivity property.

(5) The main difference between set-theoretic functions and the candidate distance function "Relative Change" ((P) in Table 4.2) is that in the former "exact match" is necessary for set operations like intersection or set difference, whereas in the case of function P it is "best match" that we will use to calculate the contribution of "attribute similarity" to the similarity between records. In our analysis, we will use both "best match" and "exact match" depending upon the attribute type/values, semantic constraints, and the analysis goals.

4.2.4 Semantic Similarity Functions - A Review

In drug-drug similarity analysis the aim is to find drugs which display similar pharmacological characteristics to the target drug. Drugs are usually coded with their FDA ("FDA", 2015) approved medical names and codes that are machine readable. These names (codes) are unique, and they are of type *categorical.* A raw comparison of any two attribute values will only result in "total dissimilarity" between drugs, although they may have "similar

Table 4.4: Generic Names of Some Drugs ("FDA", 2015)

Drug Name	Generic Name
ACANYA	ONEXTON
ACTONEL	RISEDRONATE
BACIIM	BACITRACIN
PAXIL	PAROXETINE
LIPITOR	ATORVASTATIN
NAFTINE	NAFTIFINE

indications". Hence, we need to know the semantics of drugs from the medical domain in order to determine similarity between drugs.

Example 4. *Table 4.4 shows a list of drug names and their generic brand names. Based on this semantics, we can calculate set-theoretic measures for the two sets of drugs*

$$A = \{ACANYA, BACIIM, DALMANE, KAFOCIN, LIPITOR\}$$

$$B = \{ONEXTON, ACTONEL, PAXIL, NAFTIN, ATORVASTATIN\}$$

We calculate the union, intersection, and symmetric difference for these two sets based on the "generic" semantics. The set $A \cup B$ is given below:

$$\{ACANYA, BACIIM, DALMANE, KAFOLIN, LIPITOR, ACTONOL, PAXIL, NAFTIN\}$$

$$|A \cup B| = 8$$

$$A \cap B = \{ACANYA, LIPITOR\}; \qquad |A \cap B| = 2$$

$$A\Delta B = (A\cup B)\backslash(A\cap B) = \{BACIIM, DALMANE, RAFOLIN, NAFTIN, ACTONEL, PAXIL\}$$

$$|A\Delta B| = 6$$

Substituting these values in the formulas in Equation 2 and in Table 4.3 we have the following results for the similarity of sets A and B.

$$S_{TVM}(A, B) = \frac{1}{4}, \textit{with } \alpha = \beta = 1$$

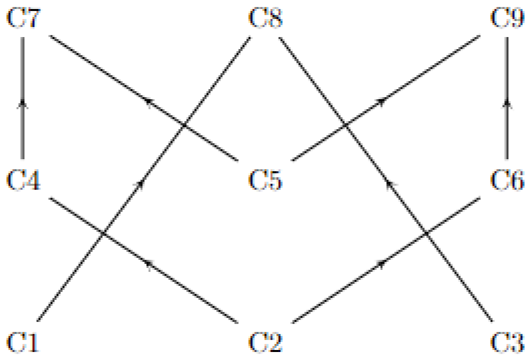

Figure 4.1: Partial Order - Hasse Diagram Example

$$JS(A,B) = \frac{1}{4} = Ham(A,B)$$

$$And(A,B) = \frac{1}{7}; \quad Dice(A,B) = \frac{2}{5}$$

The “generic” relation on the set of drugs is an equivalent relation. As such, a drug and the set of all its generic drugs are equivalent. However, semantics can be based on relations, such as “generic”.

Ontology-Supported Semantic Similarity Measures An Ontology in Healthcare domain is a collection of concept terms and their relations. Two of the well-known Ontology in Healthcare domain are SNOMED CT (El-Sappagh et al., 2018) for clinical terminology, and IC-10 ("WHO", 2015) for the classification of diseases. Concepts in an Ontology are related by *is-A* relation. We use the notation $x \preceq y$ to express the relation x *is-A* y. It means that “the concept x is *subsumed by* or *a specific class of* concept y”. That is, “concept y is more general than concept x (or subsumes x)”. An ontology structure is in general a semantic digraph (sometimes hierarchy) in which every node is an entity (concept) name and edge directed from node x to node y means $x \preceq y$. For concepts x, y, and z in an Ontology, $\preceq$ is a partial order relation satisfying the following three properties:

- *reflexivity:* for every concept x, $x \preceq x$. (Self-loops are not shown in Ontology graph structure.)
- *Antisymmetric:* for any two concepts x and y, either x and y are not related, or either $x \preceq y$ or $y \preceq x$. (Ontology structure is a directed graph.)

- *Transitivity:* For any three concepts x, y, z, if $x \preceq y$ and $y \preceq z$, then $x \preceq z$ holds. (Directed paths show transitive property.)

Thus, Ontology structure is an acyclic, directed graph. Such a graph that models a partial order is known as Hasse diagram (Graham, Knuth, & Patashnik, 1994) in Mathematics. Usually in Hasse diagram, "directions" are not shown, and assumed to be "upwards". Figure 4.1 shows such a diagram for a partial order relation on the set $\{C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8\}$ of abstract concepts. The elements in the partially set are the directed edges in the diagram. In Figure 4.1, these are

$$C_1 \preceq C_8, C_3 \preceq C_8, C_2 \preceq C_4, C_4 \preceq C_7, C_2 \preceq C_6, C_6 \preceq C_9, C_5 \preceq C_7, C_5 \preceq C_9$$

By repeatedly applying the transitive property, we compute the transitive closure that includes all derived relations. These derived relations for this example are $C_2 \preceq C_7$ and $C_2 \preceq C_9$. The observations below bring out some key properties of Ontology structure, based on which we comment and compare a few recent (Girardi et al., 2016; Harispe et al., 2014; Zhang et al., 2014) works on "Ontology-based similarity measures" for EHR analysis.

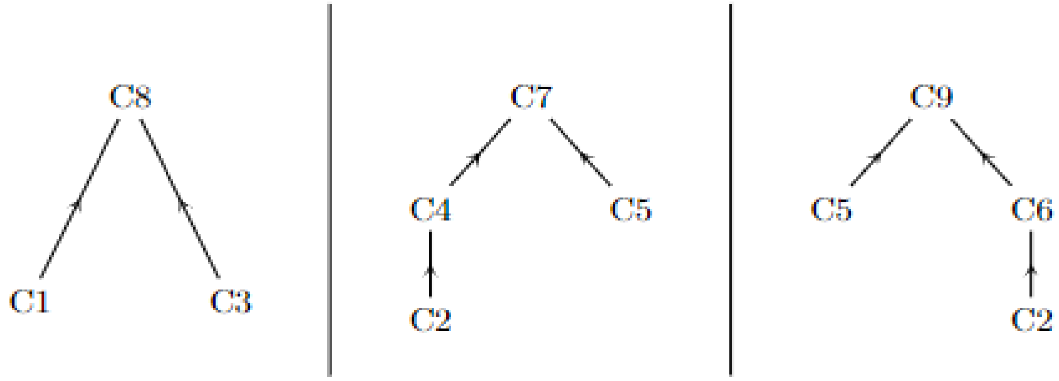

Figure 4.2: Partial Order - Rooted Hasse Diagram Example

Remark 3. *Many researchers (Girardi et al., 2016; Harispe et al., 2014; Mabotuwana et al., 2013; Zhang et al., 2014) have used similarity measures computed from the semantic distances between concept terms in an Ontology. However, after a critical analysis we found that many of the proposed functions are not precisely defined. We explain below the notation used by the above researchers, using the sample Ontology in Figure 4.1, Figure 4.2 and Figure 4.3 as reference points.*

(1) For every one of the concepts C_7, C_8 and C_9 in Figure 4.1, there is no concept subsuming it. Such concepts are called maximal, *in the sense they are "independent concepts". It is possible to split this graph into three subgraphs such that for each graph there is only one maximal concept. This is shown in Figure 4.2. Observe that the set of edges (relations) are partitioned into three sets of edges $\{C_1 \preceq C_8, C_3 \preceq C_8\}$, $\{C_2 \preceq C_4, C_4 \preceq C_7, C_5 \preceq C_7,\}$, and $\{C_5 \preceq C_9, C_2 \preceq C_6, C_6 \preceq C_9\}$. Now, each subgraph can be studied for all subsumed relations with respect to the unique "maximal" element in the graph. When the maximal element is unique, some researchers (Girardi et al., 2016; Harispe et al., 2014) call it the* root *of the Ontology.*

(2) For any two nodes x and y in an Ontology, a directed path *from x to y exists if there exists a sequence of nodes such that $< x_0 = x \preceq x_1 \preceq x_2 \preceq \cdots \preceq x_k = y >$. The length of this path is k, the number of directed edges in it. In Figure 4.1, there is no directed path from C_2 to C_5. In Figure 4.3 there is no directed path from C_4 to C_2, and there are two directed paths of lengths 2 and 3 from C_7 to C_3. A directed path from x to y ia also called a* chain *from x to y.*

(3) A shortest path *$sp(x, y)$, is a directed path from x to y such that its length is the minimum among all directed paths from x to y. In Figure 4.3, there is no path from C_4 to C_6, there is one shortest path of length 2 from C_7 to C_3, and there are two shortest paths of length 2 from C_5 to C_0.*

(4) Starting with one vertex x in an Ontology graph we can just follow the directed edges to calculate a "maximal chain". That is, we calculate $\eta =< x_0 = x \preceq x_1 \preceq x_2 \preceq \cdots \preceq x_n = y >$, and there is no element z in the graph for which $x_n \preceq z$ holds.

(5) The length of a chain η, denoted $length(\eta)$, is the number of edges in it. In Figure 4.1, the longest chains are $\eta_1 = C_2 \preceq C_4 \preceq C_7$, $\eta_2 = C_2 \preceq C_6 \preceq C_9$, $\eta_3 = C_5 \preceq C_7$, $\eta_4 = C_5 \preceq C_9$, $\eta_5 = C_1 \preceq C_8$, and $\eta_6 = C_3 \preceq C_8$, and their lengths are $length(\eta_1) = length(\eta_2) = 3$, and $length(\eta_3) = length(\eta_4) = length(\eta_5) = length(\eta_6) = 1$.

(6) Hereafter, we consider an Ontology that has a unique maximal element C_0. We call it the "root concept". From any node x in the Ontology there is at least one chain from

x to C_0. In general, there may be more than one chain from a vertex x to C_0. In Figure 4.3, from vertex C_5 to C_0 there are three chains (paths).

(7) *If $x \preceq y$, then y is called an* ancestor *of x. For a node x, the set $A(x) = \{y | x \preceq y\}$ is the set of ancestors of node x. $A(C_0) = \emptyset$, and for all other nodes x in the ontology (with a unique root concept), $A(x) \neq \emptyset$. In Figure 4.3, $A(C_5) = \{C_5, C_1, C_0, C_2, C_6, C_3\}$.*

(8) *As defined in (Girardi et al., 2016; Harispe et al., 2014; Zhang et al., 2014), the* Least Common Ancestor *of nodes x and y, denoted $LCA(x, y)$, in an Ontology is the unique concept node z in the Ontology such that it is the first "intersection of the chains" from x and y to the root node C_0. In Figure 4.3, $LCA(C_5, C_6) = C_0$ and $LCA(C_7, C_5) = C_6$. In case the Ontology is a tree, $LCA(x, y)$ is the root z of the smallest sub-tree that contains both x and y.*

(9) *The* depth *of a node x, written $depth(x)$, in the Ontology is used, but not defined in (Harispe et al., 2014) to define similarity measures. Following the definition (Daoui, Gherabi, & Marzouk, 2017), where an enhanced method to compute similarity between concept terms is discussed, $depth(x)$ is the length of the longest chain from x to C_0. In Figure 4.3, $depth(C_0) = 0$, $depth(C_1) = depth(C_2) = depth(C_3) = 1$, $depth(C_4) = depth(C_6) = 2$, $depth(C_5) = 3$, and $depth(C_7) = 4$.*

Based on the above brief discussion on Ontology structures, and the four criteria for comparing similarity measures we comment on the similarity measures used in recent patient similarity research. Many of these studies, for example in human mental health classification (Hastings, Ceusters, Jensen, K, & Smith, 2012; Larsen & Hastings, 2018) and patient-drug similarity (Daoui et al., 2017; Girardi et al., 2016; Zhang et al., 2014) depend on ontology support to define a *semantic distance function* between concept terms that occur in the EHRs. This measure, after normalization, is used to calculate the similarity of concepts. Many semantic distance measures have been compared in (Girardi et al., 2016; Harispe et al., 2014). Basically, some functions that use the "shortest path length" between the concepts, and the "path lengths of the concepts from their nearest (least common) ancestral concept" in the concept graph as its arguments are viewed as "distances" that

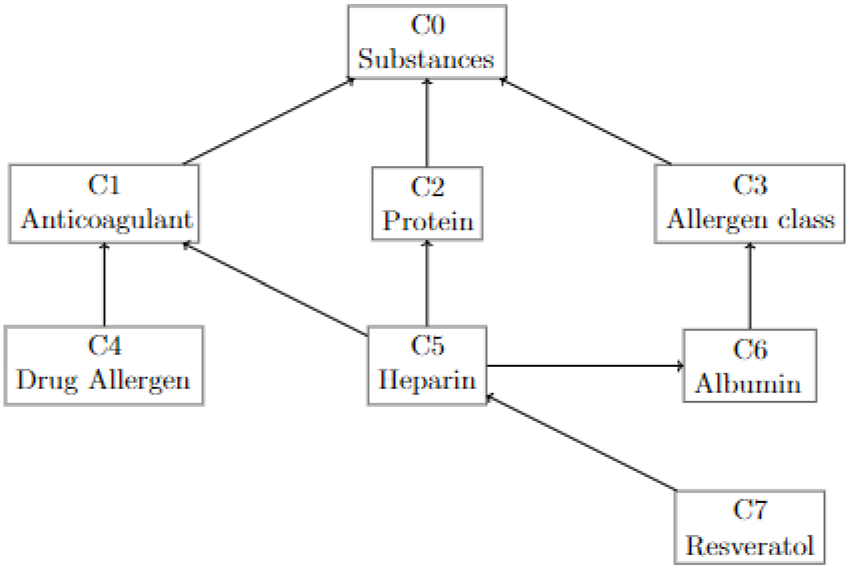

Figure 4.3: Ontology - Rooted Digraph Example

separate the concepts. However, many of the proposed functions for calculating similarity measures have flaws, as shown below.

4.2.5 Similarity Measures Based on Semantic Distance

The distance measure proposed in (Rada, H. Mili E, & Blettner, 1989) is $sp(x,y)$, the shortest distance "between x and y. Unfortunately, there may not be a directed path between every pair of concepts in an Ontology. Hence, this function cannot be used in general. Even when the Ontology is a connected digraph,several similarity measures based on $sp(x,y)$ studied in (Cordi, Lombardi, Martelli, & Mascardi, 2005; Girardi et al., 2016; Harispe et al., 2014; G. H. Wan, Wang, & Guo, 2006) involve computing "logarithms" and or "exponentiation" or "m^{th} root". Because these functions do not meet our simplicity criteria, they are not suitable for our analysis purposes. So, we consider the other similarity function in Equation 3. This function was originally proposed by Wu and Palmer (Z. Wu & Palmer, 1994) for calculating similarity of corpus terms, and is used in (Cross, 2006; Girardi et al., 2016; Harispe et al., 2014) to compute and compare similarity calculations of terms in an Ontology.

$$sim_G(x,y) = \frac{2N_3}{N_1 + N_2 + 2N_3}, \tag{3}$$

This function uses "number of nodes" (not number of edges, as is usually defined for path length) to measure chain lengths. In the definition 3, N_1 and N_2 are defined respectively as "the number of nodes from x and y to $LCA(x,y)$, and N_3 is defined as the "number of nodes on the path from $LCA(x,y)$ to the root concept C_0". The problem is that their definition of lengths do not "respect directions" and also ignore the fact that "there may be more than one path from x (y) to $LCA(x,y)$". So, the definition is "ambiguous". Moreover, it is possible that more than one chain exists from x (or y) to C_0 that are not "part of" LCA definition, and it is not clear which length should be applied to the definition of N_3.

Calculating Similarity based on Equation 3

To illustrate these ambiguous scenarios, consider the Ontology in Figure 4.3. We find $LCA(C_5, C_6) = C_0$. So, $N_3 = 1$. Because there is only one path from C_5 to C_2, $N_2 = 3$. Let N_1 be the number of nodes from C_5 to C_0. From C_5 to C_0 there are three paths, of which the path $C_5 \preceq C_1 \preceq C_0$ is outside "LCA scope". The number of nodes in this path is 3. The number of nodes in the path $C_5 \preceq C_2 \preceq C_0$ is 3 and the number of nodes in the path $C_5 \preceq C_6 \preceq C_3 \preceq C_0$ is 4. The conflict is "which value we should assign to N_1?". If we take $N_1 = 3$, we get

$$sim_G(C_5, C_6) = \frac{2}{3+3+2} = \frac{2}{8} = \frac{1}{4}$$

If we take $N_1 = 4$, we get

$$sim_G(C_5, C_6) = \frac{2}{4+3+2} = \frac{2}{9}$$

Because C_5 subsumes (specializes) C_6, intuitively they must be"close" to each other. Hence $sim_G(C_5, C_6) = \frac{1}{4}$ is more acceptable.

The functions (Harispe et al., 2014)

$$sim_H(x,y) = \frac{2depth(LCA(x,y))}{depth(x) + depth(y)} \tag{4}$$

$$sim'_H(x,y) = \frac{sp((LCA(x,y), root))}{sp(x, LCA(x,y)) + sp(y, LCA(x,y)) - sp(LCA(x,y), root)} \tag{5}$$

use depths and shortest distances. Below, through examples, we study their behavior.

Calculating Similarity based on Equation 4

This similarity function uses definition of *depth*. Let us calculate $sim_H(C_5, C_6)$. $LCA(C_5, C_6) = C_0$, and $depth(C_0) = 0$. We have $depth(C_5) = 3$ (longest), and $depth(c_6) = 2$. Substituting in Equation 4, we get

$$sim_H(C_5, C_6) = 0$$

This measure is not acceptable, because the fact that "C_5 is directly subsumed by C_6" and hence we expect them to be "closer" in similarity. We notice that $sim_H(C_4, C_5) = 0 = sim_H(C_1, C_2) = sim_H(C_2, C_3)$. However, intuitively we would like to see these pairs to be "more similar to each other. Our guess is that "depth" function (not defined in (Harispe et al., 2014)) assumes (perhaps) something different from the traditional "depth" definition in rooted Ontology. So, we avoid the suggested similarity function.

Calculating Similarity based on Equation 5

The similarity function in Equation 5 uses definition of *shortest path*. For the root node C_0, $sp(C_0, C_0) = 0$ unless the self-loop at C_0 (and every node) is admitted. But, in the rooted acyclic digraph representation, self loops are ignored in path length calculation. Consequently, this similarity function definition has the same flaw as Equation 4. So, we avoid this function.

Based on these examples, we conclude that $sim_G(x, y)$ should be appropriately redefined. A possible redefinition is that each N_i, $i = 1, 2, 3$, "denotes the number of nodes in the longest or shortest chains". With this change, we can admit it as a possible candidate for comparison with other selected candidates for similarity calculation.

4.3 A New Method for Calculating Similarity of Ontology Concept Terms Using Semantic Distances

For every concept term x in an Ontology, we can calculate *the number of nodes in a longest and shortest chain* from x to the *root* of the Ontology. Similarly, for every node x in the Ontology we can calculate *the number of nodes in a longest and shortest chain* from a leaf node to x. From these distances, we can create four different *Vector Models* for node x. These are defined below.

- *Max-Max Vector Model:*: Let $Top(x)$ denote the number of nodes in a *longest chain* from x to the *root* of the Ontology. That is, $Top(x)$ is the maximum number of concepts that subsume x. We consider all chains starting at leaf nodes of the ontology and ending at x. Among all such chains, we pick a *longest chain* and let $Bot(x)$ denote the number of nodes in that chain. $Bot(x)$ is the maximum number of concept terms that inherit x. Define the vector model of x as

  $$\langle Top(x), Bot(x) \rangle$$

  The significance of this model is that it projects every node x through the maximum number of nodes subsuming it and the maximum number of nodes subsumed by it.

- *Max-Min Vector Model:*: Let $Top(x)$ denote the number of nodes in a *longest chain* from x to the *root* of the Ontology. That is, $Top(x)$ is the maximum number of concepts that subsume x. We consider all chains starting at leaf nodes of the ontology and ending at x. Among all such chains, we pick a *smallest chain* and let $Bot(x)$ denote the number of nodes in that chain. $Bot(x)$ is the minimum number of concept terms that inherit x. Define the vector model of x as

  $$\langle Top(x), Bot(x) \rangle$$

  The significance of this model is that it projects every node x through the maximum number of nodes subsuming it and the minimum number of nodes subsumed by it.

- *Min-Max Vector Model:*: Let $Top(x)$ denote the number of nodes in a *smallest chain* from x to the *root* of the Ontology. That is, $Top(x)$ is the minimum number of concepts that subsume x. We consider all chains starting at leaf nodes of the ontology and ending at x. Among all such chains, we pick a *longest chain* and let $Bot(x)$ denote the number of nodes in that chain. $Bot(x)$ is the maximum number of concept terms that inherit x. Define the vector model of x as

  $$\langle Top(x), Bot(x) \rangle$$

The significance of this model is that it projects every node x through the minimum number of nodes subsuming it and the maximum number of nodes subsumed by it.

- *Min-Min Vector Model:*: Let $Top(x)$ denote the number of nodes in a *smallest chain* from x to the *root* of the Ontology. That is, $Top(x)$ is the minimum number of concepts that subsume x. We consider all chains starting at leaf nodes of the ontology and ending at x. Among all such chains, we pick a *smallest chain* and let $Bot(x)$ denote the number of nodes in that chain. $Bot(x)$ is the minimum number of concept terms that inherit x. Define the vector model of x as

  $$\langle Top(x), Bot(x) \rangle$$

  The significance of this model is that it projects every node x through the minimum number of nodes subsuming it and the minimum number of nodes subsumed by it.

From every node x in the Ontology the *root* concept can be reached. So, $Top(x) \geq 1$. Either x is a leaf node or it has a node inheriting it. Hence, there is a leaf node from which x can be reached. Consequently, $Bot(x) \geq 1$. So, all the vector models of every node x have positive integer component. Because we want to have bounded values, we normalize the vector by dividing its components by $Top(x) + Bot(x)$. So, the vector model for x is transformed to $\langle x_1, x_2 \rangle$, where

$$x_1 = \frac{Top(x)}{Top(x) + Bot(x)}, \qquad x_2 = \frac{Bot(x)}{Top(x) + Bot(x)} \tag{6}$$

For two concepts x and y, $x \neq y$ in the Ontology we first compute their vector models $\langle x_1, x_2 \rangle$, and $\langle y_1, y_2 \rangle$, where x_is and y_is are as defined in Equation 6. Next, we calculate their inner product. The similarity function is defined in Equation 7.

$$sim_I(x, y) = \begin{cases} 1 & \text{if } x = y \\ x_1.y_1 + x_2.y_2 & \text{otherwise} \end{cases} \tag{7}$$

It seems hard to theoretically compare these models. In Example 5 we illustrate the similarity calculation of concept terms for Max-Max model. In Section 4.7 we compare the

Table 4.5: Vector Model of Concept Terms in the Ontology in Figure 4.3

Concept::Vector Model	Concept::Vector Model
C_0:: $\langle\frac{1}{6},\frac{5}{6}\rangle$	C_4:: $\langle\frac{3}{4},\frac{1}{4}\rangle$
C_1:: $\langle\frac{2}{5},\frac{3}{5}\rangle$	C_5:: $\langle\frac{2}{3},\frac{1}{3}\rangle$
C_2:: $\langle\frac{2}{5},\frac{3}{5}\rangle$	C_6:: $\langle\frac{1}{2},\frac{1}{2}\rangle$
C_3:: $\langle\frac{1}{3},\frac{2}{3}\rangle$	C_7:: $\langle\frac{5}{6},\frac{1}{6}\rangle$

semantic similarity functions that we are proposing.

Example 5. *We consider the terms of Ontology in Figure 4.3. First, we show the steps of Max-Max vector model calculation for concept terms C_0 and C_1. Skipping similar details for other terms, we show in Table 4.5 the vector models of all concept terms of the Ontology in Figure 4.3. The inner product calculation being simple, we skip the details and show in Table 4.6 the similarity calculated by $sim_I(x, y)$ for all pairs of concept terms.*

Vector model for C_0

$Top(C_0) = 1$, because C_0 is the root and on the chain from C_0 to root there is only one node.

$Bot(C_0) = 5$, because a chain from the leaf node C_7 to root $= C_0$ is the longest, and there are 5 nodes on it.

The vector model of C_0 is $\langle\frac{1}{6},\frac{5}{6}\rangle$.

Vector model for C_1

$Top(C_1) = 2$, because C_0 is the root and on the chain from C_1 to root $= C_0$ there are two nodes.

$Bot(C_1) = 3$, because the chain from the leaf node C_7 to C_1 is the longest, and there are 3 nodes on it.

The vector model of C_1 is $\langle\frac{2}{5},\frac{3}{5}\rangle$.

Table 4.6: Max-Max Vector Model: Similarity Values for All Pairs of Concept Terms in the Ontology in Figure 4.3

	C_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_0	1	$\frac{17}{30}$	$\frac{17}{30}$	$\frac{11}{18}$	$\frac{1}{3}$	$\frac{7}{18}$	$\frac{1}{2}$	$\frac{5}{18}$
C_1	$\frac{17}{30}$	1	$\frac{13}{25}$	$\frac{8}{15}$	$\frac{9}{20}$	$\frac{7}{15}$	$\frac{1}{2}$	$\frac{13}{30}$
C_2	$\frac{17}{30}$	$\frac{13}{25}$	1	$\frac{8}{15}$	$\frac{9}{20}$	$\frac{7}{15}$	$\frac{1}{2}$	$\frac{13}{30}$
C_3	$\frac{11}{18}$	$\frac{8}{15}$	$\frac{8}{15}$	1	$\frac{5}{12}$	$\frac{4}{9}$	$\frac{1}{2}$	$\frac{7}{19}$
C_4	$\frac{1}{3}$	$\frac{9}{20}$	$\frac{9}{20}$	$\frac{5}{12}$	1	$\frac{7}{12}$	$\frac{1}{2}$	$\frac{2}{3}$
C_5	$\frac{7}{18}$	$\frac{7}{15}$	$\frac{7}{15}$	$\frac{4}{9}$	$\frac{7}{12}$	1	$\frac{1}{2}$	$\frac{11}{18}$
C_6	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{1}{2}$
C_7	$\frac{5}{18}$	$\frac{13}{30}$	$\frac{13}{30}$	$\frac{7}{18}$	$\frac{2}{3}$	$\frac{11}{8}$	$\frac{1}{2}$	1

4.4 Modifying Tversky's Method for Asymmetric Semantic Similarity Calculation for Ontology Terms

For each concept term x in the Ontology the set $A(x)$ of *features* of x can be defined (Harispe et al., 2014) as the "set of concepts subsuming it". That is,

$$A(x) = \{y | x \preceq y\}$$

is the set of *ancestors* of x. As an example, in Figure 4.3, we have $A(C_7) = \{C_7, C_5, C_1, C_2, C_0, C_6, C_3\}$, $A(C_5) = \{C_5, C_6, C_3, C_2, C_1, C_0\}$, $A(C_4) = \{C_4, C_1, C_0\}$, and $A(C_3) = \{C_3, C_0\}$. To calculate the similarity of pairs of Ontology terms (x, y), the set-theoretic similarity functions JS, AN and HA from Table 4.3 and function S_{TVM} from Equation 2 can be used. For example, the steps for calculating $sim(C_7, C_5)$ are as follows. The set $A(C_7) \cap A(C_5)$ is
$\{C_7, C_5, C_1, C_2, C_0, C_6, C_3\} \cap \{C_5, C_6, C_3, C_2, C_1, C_0\} = \{C_5, C_6, C_3, C_2, C_1, C_0\}$,
The set $[A(C_7) \cup A(C_5)$ is
$\{C_7, C_5, C_1, C_2, C_0, C_6, C_3\} \cup \{C_5, C_6, C_3, C_2, C_1, C_0\} = \{C_7, C_5, C_1, C_2, C_0, C_6, C_3\}$.

$$A(C_7) \Delta A(C_5) = (A(C_7) \cup A(C_5)) \setminus (A(C_7) \cap A(C_5)) = \{C_7\}$$

$$|A(C_7) \cap A(C_5)| = 6; \quad [|A(C_7) \cup A(C_5)| = 7; \quad [|A(C_7) \Delta A(C_5)| = 1$$

The set SS (defined in Table 4.3) for the Ontology in Figure 4.3 is the set of "all concepts in the Ontology". Thus, $|SS| = 8$. Substituting these results in the function S_{TVM} from Equation 2 and in the functions JS, AN and HA (defined in Table 4.3) we get the similarity value for the pair C_7 and C_5 shown in Equation 8.

$$sim(C_7, C_5) = \begin{cases} \frac{6}{7} & \text{for } S_{TVM} \\ \frac{6}{7} & \text{for } JS \\ \frac{3}{4} & \text{for } AN \\ \frac{7}{8} & \text{for } HA \end{cases} \tag{8}$$

If concept terms x and y are not related by partial order, then it may be acceptable to have $sim(x, y) = sin(y, x)$. That is, the symmetric property may hold for non-related concept terms. However, we think that the "symmetric" property is not compatible with the "anti-symmetric" property of partial order for concepts x and y, $x \preceq y$. To enforce anti-symmetry, we consider the set-theoretic similarity function S_{TVM} in Equation 2, proposed in (Tversky, 1977; Tversky & Krantz, 1970), after scrutinizing the properties of the parameters in the function.

For every pair of concepts x and y in an ontology, if $x \preceq y$, then $A(y) \subset A(x)$ holds. That is, the features of x includes the features of y but the converse is not true. This suggests that "the similarity of x to y is much stronger than the similarity of y to x". Hence, we need a similarity function σ which makes $\sigma(x, y) > \sigma(y, x)$. If we choose α and β in Equation 2 to satisfy the conditions $0 < \alpha < 1$, $0 < \beta < 1$, $\alpha \neq \beta$, and $\alpha + \beta = 1$, then $\sigma(x, y) \neq \sigma(y, x)$, where

$$\begin{aligned} \sigma(x, y) = S_{TVMM}(A(x), A(y)) &= \frac{|(A(x) \cap A(y))|}{|(A(x) \cap A(y))| + \alpha|(A(x) \setminus A(y)| + (1-\alpha)|(A(y) \setminus A(x))|} \\ \sigma(y, x) = S_{TVMM}(A(y), A(x)) &= \frac{|(A(x) \cap A(y))|}{|(A(x) \cap A(y))| + \alpha|(A(y) \setminus A(x)| + (1-\alpha)|(A(x) \setminus A(y))|} \end{aligned} \tag{9}$$

Let us denote $|(A(x) \cap A(y))| = a$, $|(A(x) \setminus A(y))| = b$, and $|(A(y) \setminus A(x))| = c$. Clearly,

$a = |A(y)| \geq 1$, $b \geq 1$, and $c = 0$.

$$\begin{aligned}\frac{1}{\sigma(x,y)} - \frac{1}{\sigma(y,x)} &= \Big(\frac{a + \alpha b + (1-\alpha)c}{a}\Big) - \Big(\frac{a + \alpha c + (1-\alpha)b}{a}\Big) \\ &= \frac{(2\alpha - 1)(b - c)}{a}\end{aligned} \tag{10}$$

Since $(b - c) > 0$, if we choose $0 < \alpha < 0.5$, we will have

$$\frac{1}{\sigma(x,y)} - \frac{1}{\sigma(y,x)} < 0$$

That is, $\sigma(y,x) - \sigma(x,y) < 0$, or $\sigma(x,y) > \sigma(y,x)$, which proves that "the similarity of x to y is much stronger than the similarity of y to x". By choosing α "close to 0.5" in the formulas 9 we maximize the strength of anti-symmetric property. If we apply the anti-symmetric functions (in Equation 9) to C_7 and C_5 with $\alpha = 0.4$ we get

$$\sigma(C_7, C_5) = S_{TVMM}(A(C_7), A(C_5)) = \frac{6}{6.4}$$

$$\sigma(C_5, C_7) = S_{TVMM}(A(C_5), A(C_7)) = \frac{6}{6.6}$$

Hence, $\sigma(C_7, C_5) > \sigma(C_5, C_7)$ for $C_7 \preceq C_5$. So, the set-theoretic similarity function for concepts in an Ontology is defined as

$$sim_{AS}(x,y) = \begin{cases} 1 & \text{if } x = y \\ \sigma(x,y) & \text{if } x \preceq y, \alpha = 0.45 \\ \sigma(y,x) & \text{if } x \preceq y, \alpha = 0.45 \\ JS(x,y) & \text{otherwise} \end{cases} \tag{11}$$

Example 6. *In Figure 4.3 the pairs of concepts (C_1, C_2), (C_1, C_3), (C_2, C_3), (C_1, C_6), (C_2, C_4), (C_2, C_6), (C_3, C_4), (C_4, C_5), (C_4, C_6), (C_4, C_7) are not related by the partial order $\preceq$. Hence, the symmetric function JS is used to calculate their similarity values. As an example, consider the pair (C_2, C_6). We have $A(C_2) = \{C_2, C_0\}$, and $A(C_6) = \{C_6, C_3, C_0\}$. Hence,*

$$JS(C_2, C_6) = JS(C_6, C_2) = \frac{A(C_2) \cap A(C_6)}{A(C_2) \cup A(C_6)} = \frac{1}{4}$$

The similarity of a term to itself is 1*. For the rest of the pairs of concepts, the asymmetric function* σ *is used to calculate the similarity values with* $\alpha = 0.4$*. As an example, consider the pairs* (C_2, C_5) *for which* $C_5 \preceq C_2$ *holds. We have* $A(C_5) = \{C_5, C_1, C_2, C_6, C_3, C_0\}$*, and* $A(C_2) = \{C_2, C_0\}$*,* $\mid A(C_2) \cap A(C_5) \mid = \mid A(C_2) \mid = 2$*,* $\mid A(C_2) \setminus A(C_5) \mid = 0$*, and* $\mid A(C_5) \setminus A(C_2) \mid = \mid \{C_5, C_6, C_3, C_1, C_0\} \mid = 4$*. Hence,*

$$sigma(C_5, C_2) = \frac{2}{2 + 0.4(4)} = \frac{5}{9}$$

$$sigma(C_2, C_5) = \frac{2}{2 + 0.6(4)} = \frac{5}{11}$$

Following these steps, we calculate the similarity values for all pairs of concepts in the Ontology in Figure 4.3, and show the results in Table 4.7.

Table 4.7: AS Function Similarity Values for All Pairs of Concept Terms in Figure 4.3

	C_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_0	1	$\frac{5}{8}$	$\frac{5}{8}$	$\frac{5}{8}$	$\frac{5}{11}$	$\frac{1}{4}$	$\frac{5}{11}$	$\frac{5}{23}$
C_1	$\frac{5}{7}$	1	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{10}{13}$	$\frac{5}{11}$	$\frac{1}{4}$	$\frac{2}{5}$
C_2	$\frac{5}{7}$	$\frac{1}{3}$	1	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{5}{11}$	$\frac{1}{4}$	$\frac{2}{5}$
C_3	$\frac{5}{8}$	$\frac{1}{3}$	$\frac{1}{3}$	1	$\frac{1}{4}$	$\frac{5}{11}$	$\frac{10}{13}$	$\frac{2}{5}$
C_4	$\frac{5}{9}$	$\frac{5}{6}$	$\frac{1}{4}$	$\frac{1}{4}$	1	$\frac{2}{7}$	$\frac{1}{5}$	$\frac{1}{4}$
C_5	$\frac{1}{3}$	$\frac{5}{9}$	$\frac{5}{9}$	$\frac{5}{9}$	$\frac{2}{7}$	1	$\frac{5}{7}$	$\frac{15}{16}$
C_6	$\frac{5}{9}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{5}{6}$	$\frac{1}{5}$	$\frac{5}{8}$	1	$\frac{5}{9}$
C_7	$\frac{5}{17}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{10}{11}$	$\frac{5}{8}$	1

4.5 Summary of New Results on Similarity Functions Based on Ontology Semantics

The two new results are vector models for concept terms in the Ontology, and the distance-based similarity function for these models, and asymmetric set-theoretic similarity function for terms in the Ontology.

- Distance-based similarity function for concept terms must be symmetric.

  (1) Distance-based similarity measures proposed/used by most researchers have flaws. We have proposed a new method that constructs four vector models for each concept term in the Ontology and then defines the similarity function $sim_{IN}(x, y)$

as in Equation 7. The components of each vector is normalized to have values in $[0, 1]$, and hence the inner products have values in $[0, 1]$. For achieving simplicity and efficiency we have avoided using squares and square roots for normalization. We have defined $sim_{IN}(x, x) = 1$. Hence, this function satisfies distance metric properties. The distribution of values for Max-Max vector model in Table 4.6 make us believe that the function satisfies the intuitive notion on similarity on the separation (distribution) of concept terms in the Ontology. In Section 4.7 more experimental results are shown for computing similarity between concept terms in a larger Ontology. Based on an analysis of that outcome, we will use the model(s) for similarity analysis on EHR records.

- Set-theoretic similarity functions for concept terms proposed in Equation 11 have bounded values in the range $[0, 1]$.

  (1) For concept terms x and y that are not related in the Ontology, the similarity functions S_{TVM} and JS are both symmetric and seem to be the best choices. We use JS, which is simpler.

  (2) For concept terms x and y that are related by the partial order in the Ontology, we use the similarity functions $\sigma(x, y) = S_{TVMM}(x, y)$ and $\sigma(y, x) = S_{TVMM}(y, x)$ defined in Equation 9, with the parameter value $\alpha < 0.5$. This function is anti-symmetric.

4.6 Semantic Similarity Function to Calculate Similarity Between Sets of Concepts

In this thesis our focus is in determining drug-drug similarity and use it as a basis for determining similarity between cancer patients using the EHR database. The attributes of interest are those used to model drugs, cancer diseases, and essential clinical information on the cancer patients. We use cancer disease ontology and drug ontology to treat cancer, in addition to clinical attributes of importance. The types of these attributes are either numerical or nominal (supported by ontology semantics). In an analysis we are likely to deal with sets of attributes and need to determine the similarity between two sets. Consequently, we discuss in this section methods to calculate similarity between sets of numerical values, and sets of ontology concepts.

4.6.1 Similarity Measure Between Sets of Numerical Values

In Section 4.2.2 we compared the bounded/normalized distance metrics in Table 4.2 and concluded that the relative change metric S_{RC} has the desirable properties for adapting to numeric type vector similarity calculation. Here, we use the measure $\frac{|a-b|}{max\{a,b\}}$ to calculate the "distance" between pairs of numeric values. The similarity between a and b will be

$$1 - \frac{|a-b|}{max\{a,b\}}.$$

For two sets $A = \{a_1, a_2, \cdots, a_n\}$ and $B = \{b_1, b_2, \cdots, b_m\}$, we calculate

$$s_i = \sum_{i=1}^{i=m} (1 - \frac{|a_i - b_j|}{max\{a_i, b_j\}}),$$

for $i = 1 \cdots n$. That is, s_i is the sum of similarity scores of a_i when compared with all elements of the set B. We define

$$sim(A, B) = \frac{s_1 + s_2 + \cdots + s_n}{n \times m}$$

the average of the s_is. Note that we divide by the product of sizes of the sets, because we have computed all similarity scores of every pair (a_i, b_j), $a_i \in A$, $b_j \in B$.

Example 7. *Let* $A = \{1, 3\}$ *and* $B = \{1.5, 2.6, 3.2\}$,
The similarity scores of 1 *(from set A) with respect the set B is:*

$$s_1 = (1 - \frac{0.5}{1.5}) + (1 - \frac{1.6}{2.6}) + (1 - \frac{2.2}{3.2}) = 0.667 + 0.385 + 0.312 = 1.364$$

The similarity scores of 3 *(from set A) with respect the set B is:*

$$s_1 = (1 - \frac{1.5}{3.0}) + (1 - \frac{0.4}{3.0}) + (1 - \frac{0.2}{3.2}) = 0.500 + 0.867 + 0.937 = 2.304$$

Similarity measure between A and B is the average

$$sim_{set}(A, B) = \frac{1.364 + 2.304}{6} = \frac{3.668}{6} = 0.611$$

We remark that a manual inspection of numbers in the two sets show a high similarity, which is reflected in the result.

Remark 4. *The function defined above is symmetric, That is, $sim_{set}(A, B) = sim_{set}(B, A)$. 2. The set-theoretic similarity functions compute set operations based on "exact" matches. Both "alias" and "generic drug names" qualify for "exact match". . So, if we use Equation 2 or any of its specializations in Table 4.3 we will get* 0 *as the similarity measure for Example 7. In general, the set-theoretic similarity functions can be used only for exact matching.*

4.6.2 Similarity Measure Between Sets of Concept Terms from an Ontology

The EHR of a patient will include the disease type of the patient and the set of medications prescribed for it. In general, a patient may have one or more diseases, and for each disease type, a set of medications will be assisted. We will assess patient similarity with respect to the similarity of diseases and drugs of patients. So, drug-drug similarity and disease type similarity must be first studied. Both studies require estimating the similarity of sets of concept terms from an ontology. Researchers (Cheng, Li, Ju, Peng, & Wang, 2014; Hu et al., 2017; Mathur & Dinakarpandian, 2012) have proposed methods to find disease similarity that integrate semantic and gene functional associations. The general approach is to associate each disease with a set of genes, and then calculate similarity between the gene sets using a variety of methods. For drug-drug similarity analysis, several researchers (Cordi et al., 2005; L. Huang et al., 2021; Y. Huang et al., 2019; Struckmann et al., 2021; Vilar et al., 2014; Zhang et al., 2014) have proposed methods that are specific to the type of feature sets used for drug representation. The drug models include drug codes or chemical structures or side effects for drugs. Similarity functions in such approaches need to use the "feature" specific semantics for judging "best match" to calculate the contribution of "feature set similarity" to the similarity between drugs. Many of the researchers use probabilistic estimation based on frequency of occurrence of concepts (Struckmann et al., 2021; Vilar et al., 2014), information theoretic functions involving logarithm computation (L. Huang et al., 2021), and calculating n^{th} roots of higher order functions (Cordi et al., 2005). Some have combined simpler statistical estimations, such as averaging over pairwise maximum and minimum of similarity between drugs, with information theoretic or "log likelihood score" (Cheng et al., 2014) that measures the "probability" of a functional dependence between genes. Adhering to the basic criteria that we formulated earlier, we propose below a method that is both simple and effective, and uses the ontology semantics to calculate the similarity between sets of concept terms.

In this thesis, we model a drug as a vector of features. The features include drug name, generic name, ATC codes for the drug, and sequences (strings) that inherently identify certain chemical structures of the drug. Because we will use "drug ontology" for semantics support, the similarity measures that we have already defined in Section 4.2.4 are more suitable for our analysis goal. Let $Dist_1$, $Dist_2$, $Dist_3$, and $Dist_4$ denote the four vector models used in distance-based approach proposed in Section 4.3 and let TVS_{asym} denote the asymmetric set-theoretic similarity function discussed in Section 4.4. We may use any one of the methods to compute similarity between ontology concepts in the given two sets. We can further introduce, as explained below, different statistical estimates such as maximum/minimum or averages in the calculation of set similarity. So, our proposed method is new and rich.

Method to Calculate the Similarity between sets S_1 and S_2 of concepts

Let $S_1 = \{c_1, c_2, \cdots, c_m\}$ and $S_2 = \{c'_1, c'_2, \cdots, c'_n\}$ denote two finite sets whose elements are concept terms of an Ontology. Our goal is to define a function $sim_S(S_1, S_2)$ which is bounded and symmetric. That is, we require $sim_S(S_1, S_2) = sim_S(S_2, S_1)$, and $0 \leq sim_S(S_1, S_2) \leq 1$.

- Step 1: Choose a method from the set $\{Dist_1, Dist_2, Dist_3, Dist_4, TVS_{asym}\}$, and let ρ denote the similarity function used in it to calculate the similarity between pairs of concept terms.

- Step 2: Using the function ρ, calculate the similarity value $\rho(c_i, c'_j)$ for every pair $c_i \in S_1$, $c'_j \in S_2$. That is, for each $c_i \in S_1$ we calculate the set of values

$$\sigma_i = \{\rho(c_i, c'_j) | c'_j \in S_2\}$$

Now, we have calculated the sets of values $\sigma_1, \sigma_2, \cdots, \sigma_m$, where σ_i is the set of similarity values of $c_i \in S_1$ with respect to all concepts in set S_2.

- Step 3: Calculate $\phi(\sigma_1, \sigma_2, \cdots, \sigma_m)$, where ϕ is one of the following statistics on the m sets $\sigma_1, \sigma_2, \cdots, \sigma_m$:

  (1) *MaxMax:* Calculate the maximum value Max_i of each set σ_i. That is, $Max_i = maximum\{k|k \in \sigma_i\}$. Having calculated $Max_1, Max_2, \cdots, Max_m$, define

  $$sim_S(S_1, S_2) = \phi(\sigma_1, \sigma_2, \cdots, \sigma_m) = maximum\{Max_1, Max_2, \cdots, Max_m\}$$

  (2) *MaxMin:* Calculate the minimum value Min_i of each set σ_i. That is, $Min_i = minimum\{k|k \in \sigma_i\}$. Having calculated $Min_1, Min_2, \cdots, Min_m$, define

  $$sim_S(S_1, S_2) = \phi(\sigma_1, \sigma_2, \cdots, \sigma_m) = maximum\{Min_1, Min_2, \cdots, Min_m\}$$

  (3) *MinMax:* Calculate the maximum value Max_i of each set σ_i. That is, $Max_i = maximum\{k|k \in \sigma_i\}$. Having calculated $Max_1, Max_2, \cdots, Max_m$, define

  $$sim_S(S_1, S_2) = \phi(\sigma_1, \sigma_2, \cdots, \sigma_m) = minimum\{Max_1, Max_2, \cdots, Max_m\}$$

  (4) *MinMin:* Calculate the minimum value Min_i of each set σ_i. That is, $Min_i = minimum\{k|k \in \sigma_i\}$. Having calculated $Min_1, Min_2, \cdots, Min_m$, define

  $$sim_S(S_1, S_2) = \phi(\sigma_1, \sigma_2, \cdots, \sigma_m) = minimum\{Min_1, Min_2, \cdots, Min_m\}$$

  (5) *Average:* Calculate $score_i = score(\sigma_i) = \sum_{x \in \sigma_i} x$, for $i = 1, 2, \cdots, m$. Define

  $$sim_S(S_1, S_2) = \phi(\sigma_1, \sigma_2, \cdots, \sigma_m) = \frac{score_1 + score_2 + \cdots, score_m}{n \times m}$$

Because the function is symmetric and has values in the range $[0, 1]$, the function $sim_S(S_1, S_2)$ is also symmetric and has values in the range $[0, 1]$.

Example 8. *Let $S_1 = \{C_0, C_3\}$ and $S_2 = \{C_4, C_6, C_7\}$ be two sets of concept terms taken from the Ontology in Figure 4.3.*

- *For method $Dist_1$ the similarity values for the elements of these sets are taken from Table 4.6. We have*

$$\sigma_1 = \{sim_I(C_0, C_4), sim_I(C_0, C_6), sim_I(C_0, C_7)\} = \{\frac{1}{3}, \frac{1}{2}, \frac{5}{18}\}$$

$$\sigma_2 = \{sim_I(C_3, C_4), sim_I(C_3, C_6), sim_I(C_3, C_7)\} = \{\frac{5}{12}, \frac{1}{2}, \frac{7}{19}\}$$

We have

$$Max_1 = \frac{1}{2} \quad Min_1 = \frac{5}{18} \quad Max_2 = \frac{1}{2} \quad Min_2 = \frac{7}{19}$$

Hence, the similarity measures for the two sets are as given below:

(1) MaxMax method: $sim_S(S_1, S_2) = \frac{1}{2} = 0.5.$

(2) MaxMin method: $sim_S(S_1, S_2) = \frac{7}{19} = 0.368.$

(3) MinMax method: $sim_S(S_1, S_2) = \frac{1}{2} = 0.5$

(4) MinMin method: $sim_S(S_1, S_2) = \frac{5}{18} = 0.278.$

(5) Average Method: $sim_S(S_1, S_2) = \frac{0.333+0.5+0.278+0.417+0.5+0.368}{6} = 0.399.$

- *For method TVS_{asym}, the similarity values for the elements of Sets S_1 and S_2 are taken from Table 4.7. We have*

$$\sigma_1 = \{sim_I(C_0, C_4), sim_I(C_0, C_6), sim_I(C_0, C_7)\} = \{\frac{5}{11}, \frac{5}{11}, \frac{5}{23}\}$$

$$\sigma_2 = \{sim_I(C_3, C_4), sim_I(C_3, C_6), sim_I(C_3, C_7)\} = \{\frac{1}{4}, \frac{10}{13}, \frac{2}{5}\}$$

We have

$$Max_1 = \frac{5}{11} \quad Min_1 = \frac{5}{23} \quad Max_2 = \frac{10}{13} \quad Min_2 = \frac{1}{4}$$

Hence, the similarity measures for the two sets are as given below:

(1) MaxMax method: $sim_S(S_1, S_2) = \frac{10}{13} = 0.769$

(2) MaxMin method: $sim_S(S_1, S_2) = \frac{1}{4} = 0.25$

(3) MinMax method: $sim_S(S_1, S_2) = \frac{5}{11} = 0.454$

(4) MinMin method: $sim_S(S_1, S_2) = \frac{5}{23} = 0.217$

(5) Average Method: $sim_S(S_1, S_2) = \frac{0.454+0.454+0.217+0.250+0.7609+0.400}{6} = 0.424$

Remark 5. *From our discussion above we have proposed* 20 *methods to define similarity measures between sets of concepts. These methods satisfy the original criteria set for similarity function definition. Theoretically it is hard to evaluate which method will perform well in the overall analysis. We have implemented all the methods. A sample set of results is shown in Appendix A. We compare their relative performance in many visualization charts. Our conclusion is that all our methods are "stable", in the sense they do produce measures that are fairly close to each other. A larger example is taken in the case study, and similar results are reported/compared in Appendix B and Appendix C.*

4.7 Case Study

In this section we illustrate the application of similarity function methods that we have developed so far to calculate the similarity of concept terms, and sets of concept terms arising in "mental functioning ontology". Many researchers (Babcock et al., 2021; Schriml, E, & etal;, 2018) are exploring ontology-based description to relate sub-categories of specific diseases. Just to illustrate how an ontology may have to be combined with another for similarity calculation, we have chosen the mental functioning ontology (Figure 4.4), emotion ontology (Figure 4.5), and mental disease ontology (Figure 4.6) proposed in (Larsen & Hastings, 2018).

In each ontology there are two parts. One is the core ontological types (also referred to as basic formal ontology) and the other is domain ontology to determine the different definition and different level of ontology. Mental Functioning (MF) ontology describes the patients' perceptions and behaviour, which can be observed and may not be "explicitly affective or psychiatric". Thus, it mainly contains the category concepts, such as differing the continuant and occurrent concepts. For the mental process and behaviour concepts, it lists examples of mental process concepts to better illustrate the difference. Although only a few examples are given (Larsen & Hastings, 2018), it seems to be a universal model for building other emotional ontology.

Emotion (EM) ontology has expanded the MF ontology, adding the concepts related

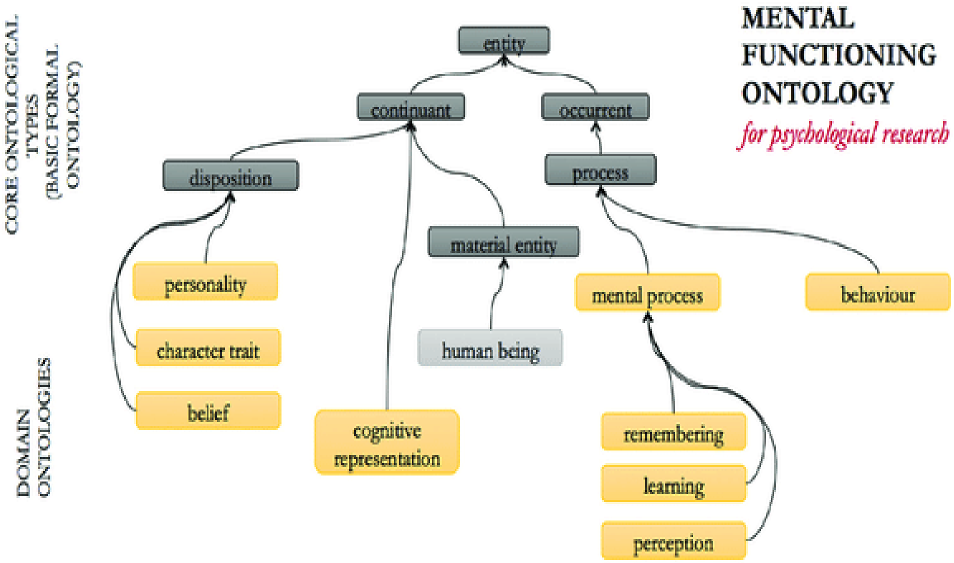

Figure 4.4: Mental Functioning Ontology

to emotions. They are often used to describe the physiological change. The term 'emotion' itself has multiple meanings and the ontology provides specific examples for declaring these ambiguities. Also, 'physiological response to emotion', 'emotion' and other high level concepts are added to the ontology. Mental Disease (MD) ontology is the aggregation of concepts describing a mental disorder. MD ontology has a different focus when expanding the ontology compared to the EM ontology, as it brings up more concepts below 'disposition' and give sub-concepts under 'thinking' and 'belief' to differ disordered concepts from its category.

After looking into the ontology, we decided to *combine* (*merge*) them into a uniform ontology to make it more suitable for comparison when studying similarities of concepts that describe emotional levels. The merging, done manually, keeps the partial order structure of MF, EM, and MD ontologies, and all their concept terms are absorbed into the new ontology, which we call *Merged Emotional Ontology* (MEO) shown in Figure 4.7. The resulting ontology is a tree with 46 concepts. Because of the tree structure there will be only one path for each node to reach the root node, thus generating 46 relationships across the ontology. In Appendix B we give the similarity measures for a small sample collection of pairs of concept terms selected from Merged Emotional Ontology, while showing the visual representation of similarity clusters for all pairs of concept terms. Given that there are

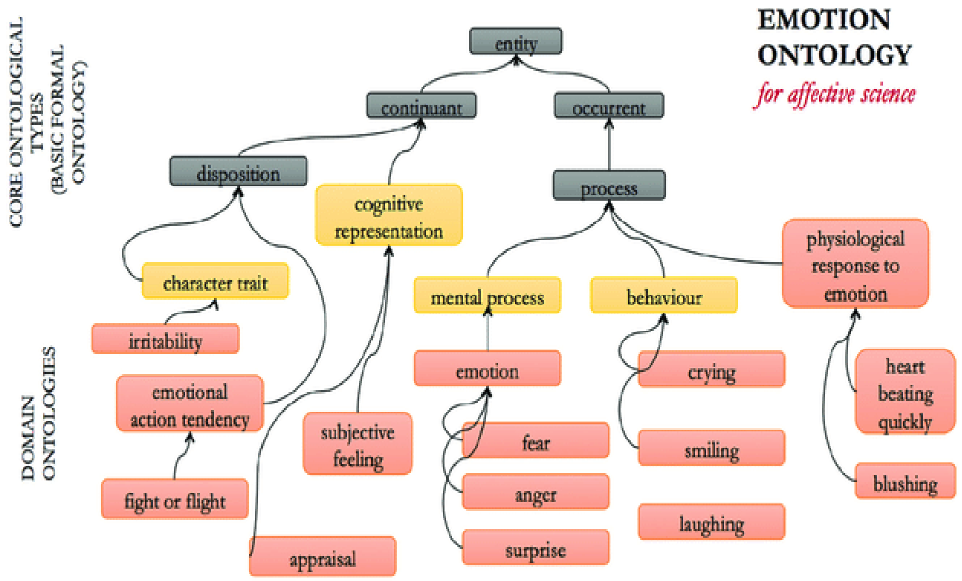

Figure 4.5: Emotional Ontology

$2^{46} - 1$ non-empty subsets of concept terms from this Ontology, we just show in Appendix C the similarity measure results for three kinds of subsets of concept terms taken from the Ontology in Figure 4.3 and the Ontology in Figure 4.7. Only an expert with knowledge from the Emotional Disease discipline can authentically validate the significance of our results. Our aim in this case study is just to illustrate the performance efficiency of our similarity function calculations, and to bring out through visualization that closely related concept terms in the ontology do have a higher similarity measure than between those terms that are far part in the ontology. This provides a convincing base on which we can justify our proposed similarity calculation methods for similarity calculations on other similar medical ontologies.

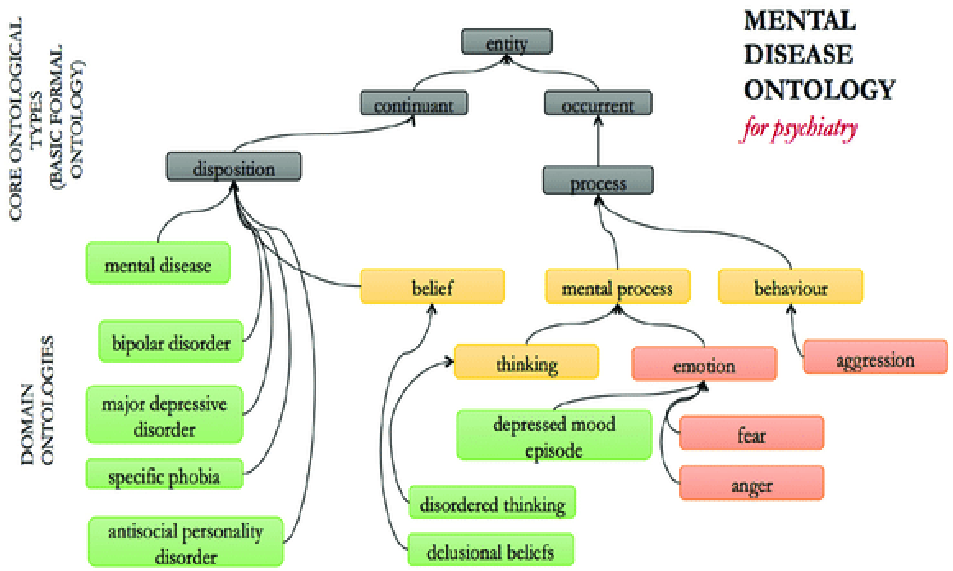

Figure 4.6: Mental Disease Ontology

4.8 User-defined Semantics

Both patient-patient and drug-drug similarity studies are conducted on data embedded in EHRs. We consider both *medical domain semantics (MDS)* and *user-level semantics* (ULS). In previous sections, we focused on MDS and selected distance-based and set-theory-based semantic similarity function candidates. In this section, we explain the ULS semantics that is relevant for EHR-based analysis. In Chapter 5 we explain how the ULS semantic constraints are integrated with our selection of similarity functions. In analyzing EHRs, the analyst can put specific attribute-level preferences for comparison and determining score for attribute matching. As part of ULS, we associate with each attribute a *mode* of matching, a *level* of significance, and a *preference* for selection. The two *modes* that we allow are "B" for *Best Match*, and "E" for *Exact Match.* Atomic values are specified in the query EHR for "E" mode. Both atomic and range (or set) of values can be specified in the query record for "B" mode. The level of significance of attributes may be set by associating weights (positive integers) in the range $[1, 5]$ with them. The weights $1, \cdots, 5$ are in increasing order of significance. The two semantic types that we allow for preferential choice of attribute values are *Lower is Better* (LB) and *More is Better* (MB) (Alagar, Alsaig, & Mohammad, 2018; Alsaig et al., 2017). These options are sufficient to express several kinds of preference combinations suggested in Alsaig et al. (2017). Table 4.8 provides the analyst's

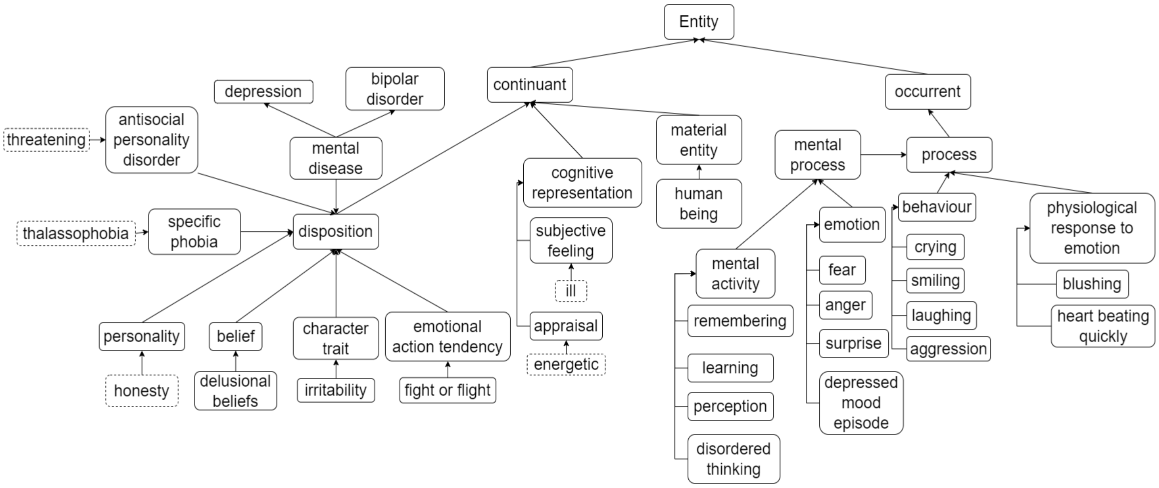

Figure 4.7: Merged Emotional Ontology

query structure. This query structure has the following assumptions and interpretations.

Table 4.8: This is an Example of User Query

Query	$[135, Medium, 5.4, \{c_1, c_2, c_3\}]$
Weight	$[4, 4, 3, 5]$
Mode	$[B, E, B, B]$
Semantics	$[LB, MB, LB, LB]$

- The Query-EHR has 4 attributes, listed in the same order as the attributes in all EHRs in the database.

- Assume that the first attribute value denotes BP level (numeric type), the second attribute denotes "exercise level" (categorical type), the third attribute denotes sugar level (numeric type), and the fourth attribute denotes medication list (set of categorical type values).

- The weights for the respective attributes are $4, 4, 3, 5$.

- The values of attributes 1, 3, and 4 are to be compared in "Best Match Mode", and value of attribute 2 is compared in "Exact Mode".

- The meaning of "semantic preference" is that the analyst prefers to select a EHR in which values of attribute 1, 3, and 4 are respectively "lower" and the value of attribute

2 is “higher” in selected records than the corresponding values specified in the Query-EHR. That is, for every record that meets this preference semantics, the similarity score will be assigned higher than the score assigned to other records that are “close” but fail this semantics.

We restrict to these attribute-level ULS constraints in our study and construct similarity functions and algorithms for similarity computation in Chapter 5.

Chapter 5

A Semantic-based Algorithm for Computing Similarity Measure of Healthcare Records

In this chapter we assume the vector model of healthcare records introduced in Chapter 3, explain the structure of an analyst's query which integrates domain level semantics (DLS) and user level semantics (ULS) of attributes in the query, and develop the algorithms for computing the similarity measure between records and a given query. A record in the healthcare database itself can be a query. In this case, the results of our algorithms reflect the relative "closeness" of healthcare records input to the algorithm.

The analyst query is one part of the analyst query structure, conforming to the vector model of the records in the dataset. That is, the query is a vector having the same number of attributes with the same order of attributes in dataset records. Every attribute of the query vector has a value from the domain of that attribute type. That is, no record or query can have "incompleteness". The semantic part of the query is another part of the query structure.

In Section 5.1 we explain the query structure. In Section 5.2 we give an overview of our algorithms that computes the similarity measure between a record and the given query. In Section 5.2 we discuss the *score functions*, which depend on the types of corresponding attributes in a record and the query. For a given pair of attributes and semantics, one scoring function is defined. For a given record-query attributes pair of a specific type, the scoring function calculates the contribution of the similarity of this pair of attributes to the similarity of that record to the query. Section 5.3 discusses the scoring functions that we need for the analyses explained in Chapter 6 and Chapter 7.

Throughout this chapter we let n denote the number of attributes in a record of the given dataset, D denote the dataset of records to be analyzed, $|D| = m$, $X =< X_1, X_2, \cdots, X_n >$ is a dataset record, and $Q =< Q_1, Q_2, \cdots, Q_n >$ is the analyst query. The attribute types are assumed to be known to the analyst. That is, the algorithm does not do any type checking.

5.1 Query Structure: Options Provided for the Analyst

The options that are provided with each attribute in the query constitute the ULS part of query. This will enable the algorithm to strictly enforce analyst's options, and not rely upon on any other external knowledge. The options are consistent with attribute types. Below, we explain the different options for different attribute type in the query.

For each attribute in the query, an analyst can specify a *Mode Option*, a *Semantic Option*, and a *weight* to indicate its criticality (significance/importance) level. Because the analyst is expected to be aware of the correct type, we assume that both X_i and Q_i are of the same type.

(1) Mode Option: The two mode options are *Exact Mode* (EM) and *Best Mode* (BM) for attribute value comparison. Only one of the options in $\{EM, BM\}$ can be specified for Q_i. By default (no option is specified), the algorithm applies EM option.

- Let the type of Q_i be *numeric*. If option EM is specified for Q_i, this value is taken to compare with the i^{th} attribute value in every record in the dataset. Strictly, the score for this comparison is either 1 (meaning equal) or 0 (meaning not equal). However, we use the function S_{RC} (Table 4.2, Chapter 4) to allow scoring values to vary in the interval $(0, 1)$. In contrast, if BM is specified for Q_i, we require a semantics to be specified for interpreting the meaning of "better/best". See below for the two kinds of semantics. Based on the specified semantics, a best score is calculated using the appropriate scoring function discussed in Section 5.3.

- Let the type of Q_i be *string* (sequence). String type attributes arise in describing target-based classification of drugs, describing protein chemical formula, specifying risk factors, and in describing environmental aspects. Due to safety considerations, only EM option is allowed for string comparison. The score for comparison of two strings is either 1 if the two strings are equal, or 0 if the strings are not equal.

- Let the type of Q_i be *nominal*. If no semantics is provided, the values are regarded as strings. As explained above, we use EM to calculate the score for the string pairs. If the semantics is given for the nominal in terms of a set of "alias (generic drug names for example)" we use BM option, and apply the Jaccard function JS (Table 4.4, Chapter 4) to calculate the score for a pair of nominal. So, the score will be a value in the interval $(0, 1)$. If an ontology-based semantics is given, then in BM option the scoring function will use one of the five methods discussed in Chapter 4 to calculate the score. The scoring functions for all these cases are given in Section 5.3.

- Let the type of Q_i be a *set type* whose elements are of type *numeric* or *string* or *nominal*. Both EM and BM options for the set above may be specified in the query. The interpretation is that the option specified (EM or BM) will be applied to every pair of set elements in score calculation. The scoring functions defined in Section 5.3 use the set similarity calculation functions discussed in Chapter 4 to calculate the score for sets. The scoring functions appropriate to EM or BM option with specified user semantics (MB and LB) are given in Section 5.3.

(2) Semantics- what is "better"?: The two semantic constraints defined for numerical type attributes in Alsaig (Alsaig et al., 2017) are *Lower is Better* (LB) and *More is Better* (MB). We use them as is for numerical type attribute comparison. That is, if numeric type Q_i is associated with LB semantics, the scoring function will assign better scoring values when Q_i is lower than X_i, and if Q_i is associated with MB semantics, the algorithm will set the lower bound for search as Q_i and assign higher scoring values whenever X_i is greater than Q_i. For nominal type attributes that are supported by Ontology we modify that definition in order to assign the minimum (for

LB case) or the maximum (for MB case) computed over the five functions defined in Chapter 4. If Q_i is not associated with any semantics, then it is treated as EM. For nominal type Q_i supported by Ontology under EM option, we use the TVS_{asym} function to calculate the score between the attribute values.

(3) Weights: We follow Alsaig (Alsaig et al., 2017) who, after an extensive investigation of investigation of weight allocation schemes for critical attributes in assessing similarity measures between service records, that the weights $\{5, 3, 1\}$ were found to be sufficient to be assigned in that order for attributes if the three levels of criminalities are *most critical*, *critical*, and *not critical*. Saaty (Saaty, 1983) gives the most general dynamic programming approach to choose weights for attributes in complex problems that require multi-criteria matching and illustrates it for real estate service domain. An easy to follow weight selection method, which is perhaps applicable to many application domains, is given in (Touran et al., 2009). The basic idea is the following:

- rank the attributes in decreasing order of importance,
- assign weighings 10 to the lowest ranked attribute, and assign weighings that are multiple of 10s in non-decreasing order to the preceding attributes in the ranked list, and
- normalize the weighings to obtain the weights.

This idea boils down to the suggestion of Alsaig (Alsaig et al., 2017) if we scale down the lowest weight from 10 to 1 and allow all attributes to have weights from the set $\{5, 3, 1\}$ of the first three odd multiples of 1. In this thesis we follow this approach. The weights are used only when the similarity measure for a record is computed, after assessing the *scores* at attribute levels.

With this background, we give a query structure in Table 5.1. The query attributes specify *disease name*, *age*, *drugs*, and *number of years since first diagnosis*. Their types are respectively *nominal*, *enumerated numeric*, *set of nominal*, and *enumerated numeric*. The analyst is made aware that *disease name* attribute has Ontology support, the *drugs* attribute has “synonyms”, and the other attributes are enumerated numerics. The analyst’s

Table 5.1: Conceptual Structure of Analyst Query

Description	Query Structure
Query	$= [\text{Bladder Cancer}, 65, \{\text{Avelumab}, \text{Cisplatin}\}, 3]$
Weights	$= [5, 3, 5, 3]$
Mode	$= [BM, BM, BM, EM]$
Semantics	$= [MB, LB, *, MB]$

query has specified values for these attributes, the modes for search, the semantics for scores, and the weights for attributes. The meaning of these specifications for attribute *drugs* is as follows:

- *weight:* This attribute is critical, and hence assigned the highest weight.
- *Mode and Semantics:* They are taken together to decide on search and similarity calculation options. The search option required by the analyst is BM under MB. So, Ontology is used to compare cancer disease names with "Bladder Cancer" and the similarity measures are calculated using all the five methods given in Chapter 4. Because of MB semantics, the maximum of the five similarity measures is assigned as score for the pair (Blood Cancer, r), where r is the value in the database record against which the current comparison is done.

Similar interpretations can be done for other attribute values in the query.

5.2 General Overview of the Ranking Algorithm

The analyst query structure, as in Table 5.1, is received by the algorithm, and is preprocessed for similarity computation ranking. From the input query, the preprocessor extracts the set of attributes of the query and then projects all EHRs in the hospital dataset on these attributes, while retaining a unique pseudo identifier Pid for each projected record. The Pid of each projected is linked to the corresponding Patient Identifier PID of the record. So, the privacy of patient will not be compromised, and after ranking the ranked records can be linked to their corresponding PIDs in order to perform additional analysis that might be required. Let W be the column matrix of weights for attributes. That is,

$$W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

Let $D = \{R_1, R_2, \cdots, R_m\}$, be the collection of projected records. We can view each record R_i as a vector $R_i =< X_{i1}, X_{i2}, \cdots, X_{i,n} >$, in which the attributes and their types are identical to the analyst's query $Q =< Q_1, Q_2, \cdots, Q_n >$. The algorithmic steps for comparison, matching, calculating scores for attribute pairs, and calculating similarity between Q and each R_i are as follows.

(1) *Compare Q with R_i:* This comparison is done pairwise $Q_j :: Xij$, for $1 \leq i \leq n$, by considering the mode, range, semantic options specified in the query structure. The scoring functions used to calculate $\sigma_{ij} = score(Q_i, X_{ji})$ are explained in Section 5.3. At this stage we only need to emphasize that σ_{ij} is the similarity score resulting from the comparison of the j^{th} attribute Q_j of the query with the j^{th} attribute X_{ij} of the i^{th} record R_i. The algorithm ensures that σ_{ij}s are normalized.

(2) *Repeat for all Records:* Do Step 1 for $i = 1, \cdots m$.

(3) *Form the matrix of scores:* Put the results of the first two steps in matrix AS of attribute scores. The rows represent the records and the columns represent the attributes.

$$AS = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{mn} \end{bmatrix}$$

(4) *Calculate the weighted sum of scores for each record:* We calculate the product matrix $WS = AS \times W$. That is, the weighted similarity measure for i^{th} record is $WS[i]$ as given below.

$$WS[i] = w_1\sigma_{i1} + w_2\sigma_{i2} + \cdots + w_n\sigma_{in}$$

(5) *List of Similarity measures:* The list $WS[1], WS[2], \cdots, WS[m]$ contains the final similarity measures for the m records.

(6) *Rank the records:* Sort this list in decreasing order of similarity measures. Link each element of the sorted to the Pid, which is linked to PID, and get the ranked records.

This ranked patient list can be used for further analysis of EHRs.

5.3 Algorithm Details: Scoring Functions

In this section we define scoring functions for different attribute types. All scoring functions integrate the ULS with the similarity functions defined in Chapter 4. We use the notation r and q to respectively denote the values of record and query attributes under discussion.

5.3.1 Scoring Function for Numeric Type Attribute Pair

We consider the options EM, BM with LB semantics, and BM with MB semantics.

- *Exact Mode (EM):* In Chapter 4 we selected the distance function S_{RC} (defined in Table 4.2) as most appropriate to define similarity as $1-S_{RC}$. So, the scoring function that we use is as shown in Equation 12

$$score(r,q) = \begin{cases} 1 & r = q \\ 1 - \frac{|r-q|}{max(r,q)} & r \neq q \end{cases} \tag{12}$$

  This function is symmetric, and normalized to have values in $[0,1]$.

- *Best Match (BM) with MB Semantics:* We want to "reward" the cases where $r > q$ and "penalize" the cases where $r < q$. If $r > q$ for a record then, $max(r,q) = r$, and if $r < q$ in a record then $max(r,q) = q$. If $r = q$ in a record, then we assign 1 to $score(r,q)$. For MB semantics, consider the two cases:

  ○ case $r > q$ We have $max(r,q) = r$ and want $score(r,q)$ to be greater than 1. We achieve this when we add the "relative change" $\frac{|r-q|}{max(r,q)} = \frac{|r-q|}{r}$ to 1, and assign it to $score(r,q)$.

- ○ case $r < q$ We have $max(r, q) = q$ and want $score(r, q)$ to be less than 1. We achieve this when we subtract the "relative change" $\frac{|r-q|}{max(r,q)} = \frac{|r-q|}{q}$ from 1, and assign it to $score(r, q)$.

Thus, the score function is as defined in Equation 13.

$$score(r, q) = \begin{cases} 1 & r = q \\ |1 - \frac{|r-q|}{q}|, & r < q \\ |1 + \frac{|r-q|}{r}|, & r > q \end{cases} \tag{13}$$

- *Best Match (BM) with LB Semantics:* We want to "reward" the cases where $r < q$ and "penalize" the cases where $r > q$. Hence, we just reverse *score* functions defined for MB semantics. The scoring function for LB semantics is as shown in Equation 14.

$$score(r, q) = \begin{cases} 1 & r = q \\ |1 + \frac{|r-q|}{q}|, & r < q \\ |1 - \frac{|r-q|}{r}|, & r > q \end{cases} \tag{14}$$

5.3.2 Scoring Function for String (Sequence) Attribute Pair

In healthcare domain string (sequence) type attributes arise to describe drug names, protein sequencing, and for recording chemical formulas. In all these cases, for the sake of safety only "exact match" (EM) is consider. So, for two strings r and q the scoring function is defined as in Equation 15.

$$score(r, q) = \begin{cases} 1 & r = q \\ 0 & r \neq q \end{cases} \tag{15}$$

5.3.3 Scoring Function for Nominal Attribute Pair

In healthcare domain string nominal type attributes arise to describe brand names, generic names or synonyms of drugs.

- *No ontology support exists:* Only "exact match" (EM) is considered, wherein a brand name may be treated as "equivalent to" its generic name. So, for two nominal r and q the scoring function is defined as in Equation 16.

$$score(r, q) = \begin{cases} 1 & r = q \ or \ q \in \text{ Synonym}(r) \\ 0 & r \neq q \ or \ q \notin \text{ Synonym}(r) \end{cases} \tag{16}$$

- *Ontology Support exists - EM:* The similarity measure TVS_{asym}, defined in Chapter 4 is computed for the pair of terms r and q. The scoring function is given in Equation 17.

$$score(r, q) = TVS_{asym}(r, q) \qquad \text{(See Chapter 4)} \tag{17}$$

- *Ontology Support exists - BM:* If no semantics is given, the scoring function in Equation 17 is used. Otherwise, we calculate the 5 similarity measures $Dist_1$, $Dist_2$, $Dist_3$, $Dist_4$, and TVS_{asym} defined in Chapter 4 for the pair (r, q). Under MB semantics, $score(r, q)$ is assigned the maximum of the five computed measures for the pair (r, q). Under LB semantics, $score(r, q)$ is assigned the minimum of the five computed measures for the pair (r, q). Thus, the scoring function is as defined in Equation 18.

$$score(r, q) = \begin{cases} max\{\{Dist_i(r, q)\}_{i=1}^4, TVS_{asym}(r, q)\} & \text{(MB)} \\ min\{\{Dist_i(r, q)\}_{i=1}^4, TVS_{asym}(r, q)\} & \text{(LB)} \end{cases} \tag{18}$$

5.3.4 Scoring Function for Sets

In healthcare domain many attributes such as blood pressure reading, glucose reading, and dosage of drugs are numeric types. While comparing patient records, we may have to compare the readings over a period (may be over several visits) for two patient records. That is, we need to consider two sets of numeric values and estimate its similarity. Similarly, we may have to assess the similarity between two sets of strings, and two sets of nominal. We first calculate pairwise $score(x_i, y_j)$, $x_i \in S_1, y_j \in S_2$, using one of the score functions defined above, and next use the "Average" method given in Chapter 4 to calculate $score(S_1, S_2)$. Hence, the general function is as given Equation 21.

$$score(S_1, S_2) = \begin{cases} 1 & S_1 = S_2 \\ \frac{\sum_{x_i \in S_1} \sum_{y_j \in S_2} score(x_i, y_j)}{|S_1||S_2|} & S - 1 \neq S_2 \end{cases} \tag{19}$$

We have the following list of different attribute-level score functions.

- *Set of Numeric Type:*

  EM option: The score for every pair of elements is calculated using the function in Equation 12.

  BM Option, MB Semantics: The score for every pair of elements is calculated using the function in Equation 13.

  BM Option, LB Semantics: The score for every pair of elements is calculated using the function in Equation 14.

- *Set of String Type:* Only EM option is allowed. The score for every pair of elements is calculated using the function in Equation 15.

- *Set of Nominal Type - no ontology - synonyms* Only EM option is allowed. The score for every pair of elements is calculated using the function in Equation 16.

- *Set of Nominal Type - ontology exists*

  EM option: The score for every pair of elements is calculated using the function in Equation 17.

  BM Option, MB and LB Semantics: The score for every pair of elements is calculated using the function in Equation 18.

Scoring functions for pairs of fields of EHR records whose types are of the same Record type can be calculated by applying the general algorithm in Section 5.2, assuming the weights of attributes are 1. For fields whose types are sequences, if their types are equal then the scoring function can be applied pairwise to the elements of the two sequences. The result will be a sequence of score values. Statistical measures, such as median and mean, as necessary may be computed on the sequence.

Chapter 6

Drug-Drug Similarity Calculation

In this chapter we study drug-drug similarity for cancer domain. We use the scoring functions that we developed in Chapter 5 to calculate drug-drug similarity measures. Two kinds of experiments are conducted. One experiment calculates drug-drug similarity of every pair of records in a drug database, and with respect to each record the other records are ranked in decreasing order of their similarity measures. In the second experiment, for a given query structure (can be a drug record with analyst preferences and semantics), we produce a ranked list of drug records in decreasing order of similarity. The results of these experiments may be used by experts/analysts for their investigation, such as *similarity of side effects of similar drugs*, *clustering of drugs*, and *interference or interaction relationship of drugs used by cancer patients who have other diseases such as diabetics or allergy.*

6.1 Related Work

Drugs play a significant role for humans to overcome diseases, bring a level of safety and improve quality of life. Studying the comparative effects of drugs is hard by just using clinical data. Researchers have been studying the chemical structures, side effects, and drug-drug interactions to help improve patient health and help drug manufacturers in their costly and time consuming research. It is in this context, they started using computational methods and analyzing drug-drug and drug-patient interactions.

Computational methods for studying drugs, their effects on patients, and their interactive behaviour include calculating similarity measures (Ferdousi et al., 2017; L. Huang et al., 2021; Struckmann et al., 2021; Vilar et al., 2014; Zhang et al., 2014). These methods are based on different hypotheses, as enumerated below.

- *Protein Sequence:* In (Zhang et al., 2014), a drug model is its protein sequences (*"DRUGBANK" online*, 2017). Similarity between drugs is measured using Smith-Waterman alignment score (Okada, Ino, & Hagihara, 2015). The patient taking a drug is modeled by the set of ICD9 codes of drugs. Similarity between patients is calculated using Jaccard function (see Chapter 4). The paper gives results of their experiments on real world datasets.

- *3D Macophoric Approach:* Drug-drug interaction is studied in (Vilar et al., 2014) using "macophoric approach", a procedure used for finger print matching. This method maps a drug into a vector of integer codes. After mapping a drug to a vector model (of 0s and 1s) they use the modified Jaccard formula

  $$Sim(A,B) = \frac{N(A\&B)}{N(A) + N(B) - N(A\&B)},$$

  where $N(A)$ and $N(B)$ respectively denotes the number of features present in structure A and B, and $N(A\&B)$ is the number of features common to both structures A and B. The methods they use to assess the performance of the model are very specific to the characteristics of 2-D and 3-D molecular structures. It is "domain and expert-centric" and is beyond the comprehension of software analyst. Because their approach is so unique to the biomedical discipline the paper does not provide any comparison of their method with others.

- *Based on Biological Elements:* Drug-drug interaction is studied in (Ferdousi et al., 2017) using the five biological elements *Carriers*, *Transporters*, *Enzynme*, and *Targets* (CTET) as the basis for a drug model. Each biological element is encoded as a binary vector. Using the Drug Bank (*"DRUGBANK" online*, 2017), the experts (research analysts in biomedical domain) found out that there are 23 distinct carriers between drugs. So, they set the length of Carrier Vector to 23. The value of i^{th} vector component was set to 1 (a positive value) through which "the corresponding carrier was in association with related carrier". Otherwise, it was set to 0. Similarly, they set the lengths and values for other vectors. Transport Vector of length 115, Enzyme Vector of length 235, and Target Vector of length 1787 are also binary vectors constructed in a similar manner. So, the CTET vector of length 2004 is constructed for each drug. They have compared 12 different similarity measures, some distance-based, some inner product-based, and some "correlation-based", to calculate the drug-drug similarity of more than $45,000$ drug pairs. They have used the "Russell- Rao" correlation measure $\frac{N(A\&B)}{L}$, where $N(A\&B)$ is the number of 1s common to both CTET vectors A and B, and L is the length of the CTET vector. Their primary conclusions are the following:

(1) For many drugs the "biological elements" are not found in Drug Bank (*"DRUGBANK" online*, 2017). Hence, their findings are incomplete.

(2) A threshold $\delta > 0$ is fixed and only those drug pairs whose similarity measure exceeds δ are selected to have exhibited "significant interaction".

(3) Because of the above two reasons, the DDI for about 28% of drugs cannot be observed.

(4) There is also some "false positive" effect. They remark that "the higher similarity does not necessarily directly refer to higher severity of DDI". They attribute the reason "may be one biological element has strong influence (because of the large number of overlaps of 1s) which tilts the balance in the overall computation of similarity measure".

- *Review of Methods:* In this article (Y. Huang et al., 2019), many methods for assessing drug-drug similarity are reviewed. These are *ATC-based Similarity Method* (ATCM), *Chemical Structure-based Similarity Method* (CSSM), *Targets-based Similarity Method* (TBSM), *Drug Interaction-based Similarity Method* (DISM), and *Side Effect-based Similarity Method* (SESM). These methods differ in their "similarity hypotheses", their "modeling" of drug information, and in their "choice of similarity functions". For example, the similarity hypothesis of ATCM is "two drugs are similar if they have similar ATC codes (Sketris, Metge, Ross, & MacCara, 2004)", whereas the similarity concept of TBSM is "two drugs are similar if their biological targets (proteins) have similar structure". Many researchers use *information theoretic* functions for ATCM, whereas for TBSM they use Jaccard score (L. Huang et al., 2021). A full review of these methods reveal that not one solution fits all, and hence the difficulty in comparing the efficiency or performance of similarity-based methods.

In summary, we observe the following on the reviewed methods.

(1) Most of the methods use the Jaccard formula (see Chapter 4) or one of its variations. Jaccard formula uses only "exact match" of set elements to compute set intersection and set union, unless "generic or alias" is defined for set elements. That is, the measure does not include any semantics, except for the domain semantics used in drug encoding.

(2) Drug model is different for different analysis goals. There is no work on evaluating sufficient completeness of any model. Also, for some drugs certain code types are not yet available.

(3) The CTET vector model (Ferdousi et al., 2017) uses the set-theoretic Jaccard function, although their drug model is a "vector". As observed by Tversky (Tversky, 1977) this is not appropriate, and will lead to erroneous conclusion, because it is possible to have three binary vectors X, Y, Z of equal length L for which $\frac{N(X\&Z)}{L} = \frac{N(Y\&Z)}{L}$, although $X \neq Y$.

(4) Ontology-based similarity functions have not been used by most of the researchers.

(5) None of these methods have considered user-centric semantics and preferences (stating weights to discriminate the importance of one biological element over another) for calculating drug-drug similarity.

6.2 Motivation

Patient-patient similarity plays an important role in personalized medicine (Y. Huang et al., 2019; J. Lee et al., 2015; Parimbelli, Marini, Sacchi, & Bellazi, 2018; Sharafoddimi, Dublin, & Lee, 2017; Wang et al., 2017; Zhang et al., 2014). As reviewed above, the hypothesis "patients P_1 and P_2 are similar if they take similar drugs" was first studied in (Zhang et al., 2014). We are inspired by this hypothesis, and pursue it in this thesis by offering a completely different approach to model drugs and model patients. Further, wy reckon that "by relating drug-drug similarity to patient-patient similarity, we are may enable targeted medical prescription to only the likely cohort who are clinically similar". Motivated by these factors, we investigate drug-drug similarity in this chapter, and use it to study patient-patient similarity in Chapter 7.

We restrict to studying cancer drugs administered to cancer patients. The rationale for this restriction is two fold. First, cancer disease is affecting people of all ages and ethnicity. It is one of the most severe diseases and perhaps most expensive to cure. So, our contribution might benefit a large worldwide community of patients and researchers. Second, studying drug-drug similarity on drugs restricted to one "disease" might give better results than studying it on all types of drugs. The hypothesis is "drugs for one disease are more related". There are several types of cancer and hence there are many types of drugs. So in our research, we mainly focus on the drugs being used to deal with cancer type diseases. As cancer type diseases are mainly caused by malfunctioning cells, the drugs proscribed to them may have higher similarity than other drugs.

6.3 Drug Model

All researchers whose work have been reviewed in Section 6.1 have used the domain-specific coding to represent drug vector as a 0/1 vector. Most of them have used only one attribute to code in the vector model. The exception is the "biological model" (Ferdousi et al., 2017) in which four different attributes were coded. Because the vector model of drug is taken as a *concatenation* of the four coded sub-vectors, and they have chosen an arbitrary ordering for them, their representation is ambiguous. More importantly, the coded 0/1 vector is not amenable to "ontology-based" semantic support. That is, assuming that an attribute (say, Enzyme) has an ontology, then two different enzymes in the ontology have a "semantic relationship", which is lost in the two coded strings of these two attribute values. As opposed to these models, our drug model is richer, includes five different attributes that are "related", and many of the attributes have ontology-based semantic support. The attributes that we chosen and their types are shown in Table 6.1. Considering their long R&D cycle, it's common that each specific drug have many attributes related to them. To find the best attributes describing the characteristic of a drug and also make it possible for us to calculate the similarity between drugs, we have selected the attributes that we need for our calculation. We use the semantic similarity functions that we have developed in Chapter 4 to assess the score between attribute values whenever the attribute semantics is supported by an ontology.

Table 6.1: Drug-Drug Similarity Attributes

Attribute Name	**Type**	**Example Value**
Generic Name	Nominal	Gemcitabine
Brand Name	Sets of Nominal	Gemzar, Infugem
ATC Codes	Nominal	L01BC05
Cancer Names	Sets of Nominal	lung non-small cell carcinoma,cervical cancer
Dosage Strength (mg)	Numerical	100
Drug Side Effect	Set of String	increased bleeding increased Thrombosis increased infection

6.3.1 Attributes and Their Types

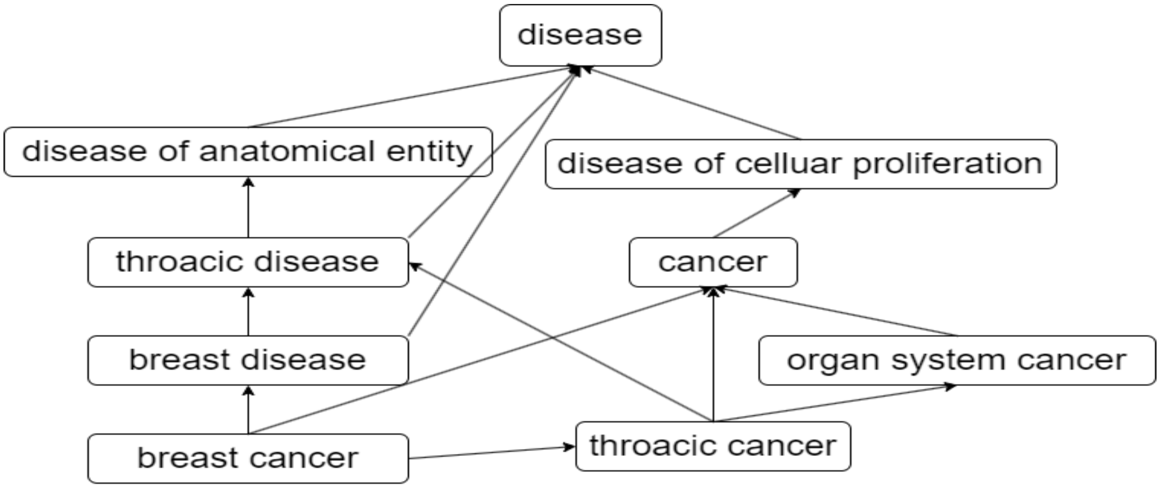

Figure 6.1: Disease Ontology - Breast Cancer Example

Drug Names: Often a drug has multiple names related to it, such as a generic name, and a brand name. We choose to include the generic name as the identification of a certain drug, and the brand name to be used for further comparison. if two drugs have identical names, the scoring function for “Brand Name” comparison will return 1.

Cancer Names: Our goal mainly is focused on the drugs proscribed to patients with cancer diseases. Cancer names are a set of nominal type values, showing the disease which the drug is targeting. A drug may target only one type of cancer or may target multiple cancer types at the same time. The cancer disease types categorized in the ontology (T. J. Wu & et al;, 2015) is very simple and does not include staggered relationships among different cancer diseases. So, we use the disease type ontology (Whetzel & et al;, 2011) proposed by BioPortal which includes cancer diseases in the ontology. In this ontology, we find all the possible diseases included. For each disease in the ontology, we can extract a small portion of the whole ontology to show only the nodes that are related for a specific study. The full cancer ontology, which is part of the Bioportal ontology (Whetzel & et al;, 2011) is too large and hard to show. For breast cancer, we extracted the ontology shown in Figure 6.1 from the full ontology (Whetzel & et al;, 2011). The extracted part shows the breast cancer as the leaf node, and it includes the top-level disease categories such as 'breast diseases' and 'disease of anatomical entity. In Appendix D we explain the different ontology extracted from Bioportal for our experiments.

ATC Codes : We use the ATC codes (Sketris et al., 2004), the classification system followed in Canada and is maintained by the WHO. It classifies the active ingredients of drugs. The attribute type is nominal. There exists an ontology, shown in Figure D.1 (Appendix D) for the codes, which fits well into our research. Each ATC code represents only one single drug and all the drugs only have one ATC code.

Dosage Strength: Every drug in the market is associated with a "dosage". In personalized medicine, dosage is an important attribute for self-administration of drug by a patients. The type of Dosage is numeric and the dosages of all drugs with a brand name are assumed to have the same unit of measurement. Because a brand name associated with a dosage must be regarded different with the same brand name associated with a different dosage, in the drug database more than one drug record will exist for a brand name, This helps to identify drugs with patients in personalized medication.

Drug Side Effect: A drug name associated with its dosage may have specific effects. To simplify the representation, we list key words or a simple string of keywords to describe side effects.

Of these 6 attributes we the 5 attributes 'Brand Name', 'ATC Codes', 'Cancer Names',

Table 6.2: Drug Record Scoring Functions

Attribute	Type	Mode	Scoring Function
Brand Name	Nominal	EM	Equation 16
ATC Codes	Ontology	BM	For LB and MB: Equation 18
Cancer Name	Ontology	EM or BM	EM: Equation 20 BM: Equation 18
Dosage	Numerical	EM or BM	EM: Equation 12 BM: LB: Equation 14 BM: MB: Equation 13
Drug Side Effect	Set of Nominal	EM	Equation 17

’Dosage Strength’, and ’Drug Interaction’ for comparison and similarity assessment. We assign different weights (in the range $1 \cdots 5$) to them to discriminate their level of significance.

6.4 Drug-Drug Similarity Calculation

For our Drug-Drug similarity comparison, we have collected 50 drug records from Drug Bank (*“DRUGBANK” online*, 2017; "FDA", 2015). The records are listed in Appendix D. Each drug record will contain values for all the 6 attributes we mentioned before. For the attributes ’Brand Name’, ’Drug Side Effect’ whose types are ’Set of Nominal’ and ’Set of String’ only “EM"" mode is used for comparison, because the values are atomic not supported by any semantics. The attributes ’ATC Code’ and ’Cancer Names’ have ontology support for comparison. So, we can use “BM” match. The type of attribute ’Dosage Strength’ is numeric. So, the analyst can choose either ’EM’ mode or ’BM’ mode for comparison. For each mode and semantics (LB/MB) scoring functions defined in Chapter 5 are selected to calculate matching scores between corresponding attribute values in records. For our selection of drug record attributes, the scoring functions (their references to Chapter 5) are shown Table 6.2. The full algorithm described in that chapter is used to calculate the weighted average of scores.

First, we implemented the method for the 5 drug records shown in Table 6.3 for both modes EM and BM and calculated the similarity between all pairs of drug records. Next, we expanded the database size to 10 drugs and repeated the calculations. The results of this experiment does not show any significant improvement. So, we next repeated the experiment for all 50 drugs. The detailed results and the drug database used in the full experiment on 50 drugs are given in Appendix D.

Table 6.3: Drug Record

Drug Name	**Brand Name**	**ATC Code**	**Cancer Names**	**Dosage (mg)**	**Drug Side Effect**
Gemcitabine	Gemzar, Infugem	L01BC05	lung non-small cell carcinoma, invasive bladder transitional cell carcinoma, cervical cancer, head and neck carcinoma, lung small cell carcinoma, breast cancer	100	increased bleeding, increased infection, increased Thrombosis
Porfimer sodium	Photofrin	L01XD01	Esophageal Cancer, lung non-small cell carcinoma	2.5	increased Thrombosis, increased photosensitizing
Gefitinib	Iressa	L01XE02	lung non-small cell carcinoma	250	increased Thrombosis
Etoposide	Etopophos, Toposar, Vepesid	L01CB01	Merkel cell carcinoma, lung non-small cell carcinoma, ovarian cancer, prostate cancer, retinoblastoma, thymic carcinoma, testis refractory cancer	20	increased bleeding, increased Thrombosis, increased infection
Tamoxifen	Nolvadex-D, Soltamox	L02BA01	breast cancer, estrogen-receptor positive breast cancer, ovarian cancer	10	increased bleeding, increased Thrombosis

6.4.1 Experimental Results - Exact Match Mode

Under Exact Match (EM) mode, the only thing we need to care about is to find the "nearest values" in target records for each attribute of the query record. Under this semantic, we can calculate the similarity matrix for all the drugs. We first resolved one difficulty. In the comparison of the attribute values for 'Cancer Names', what we would like to see depends on the semantic distance (semantic relation) of diseases rather than levels of depth". So, we chose TVS_{asym} function for calculating the similarity in *Cancer Names*. But this function is asymmetric. That is, we will get

$$Sim_{AB} = TVS_{asym}(A, B) \neq TVS_{asym}(B, A) = Sim_{BA}$$

For different drug names $Drug_A$ and $Drug_B$ we want to see "equal" similarity measure, regardless of the "direction of measurement" (whether $Drug_A$ is compared with $Drug_B$ or vice versa). To solve this problem, we have defined

$$Sim_{AB} = Sim_{BA} = \max\{TVS_{asym}(A,B), TVS_{asym}(B,A)\} \tag{20}$$

We use the weight $[4,2,2,1,2]$ for the 5 attributes to calculate the similarity measures, and we standardize all the similarity values into the range $(0,1)$, making it possible to visualize all the results we get from the algorithm. The results of the resulting method are shown in Figure 6.2. We remark that the drug 'Etoposide' and 'Gemcitabine' have "high similarity". This is fair since they have identical potential *DrugInteraction* and both can be used for 'lung non-small cell carcinoma'. Drugs 'Tamoxifen' and 'Gefitinib' have "low similarity". This is acceptable since they only share few ATC code nodes, and both are focused only on a few cancer types. Given these result on a small sample of records, we can claim that the method that we use has the potential to turn in valid results on large datasets.

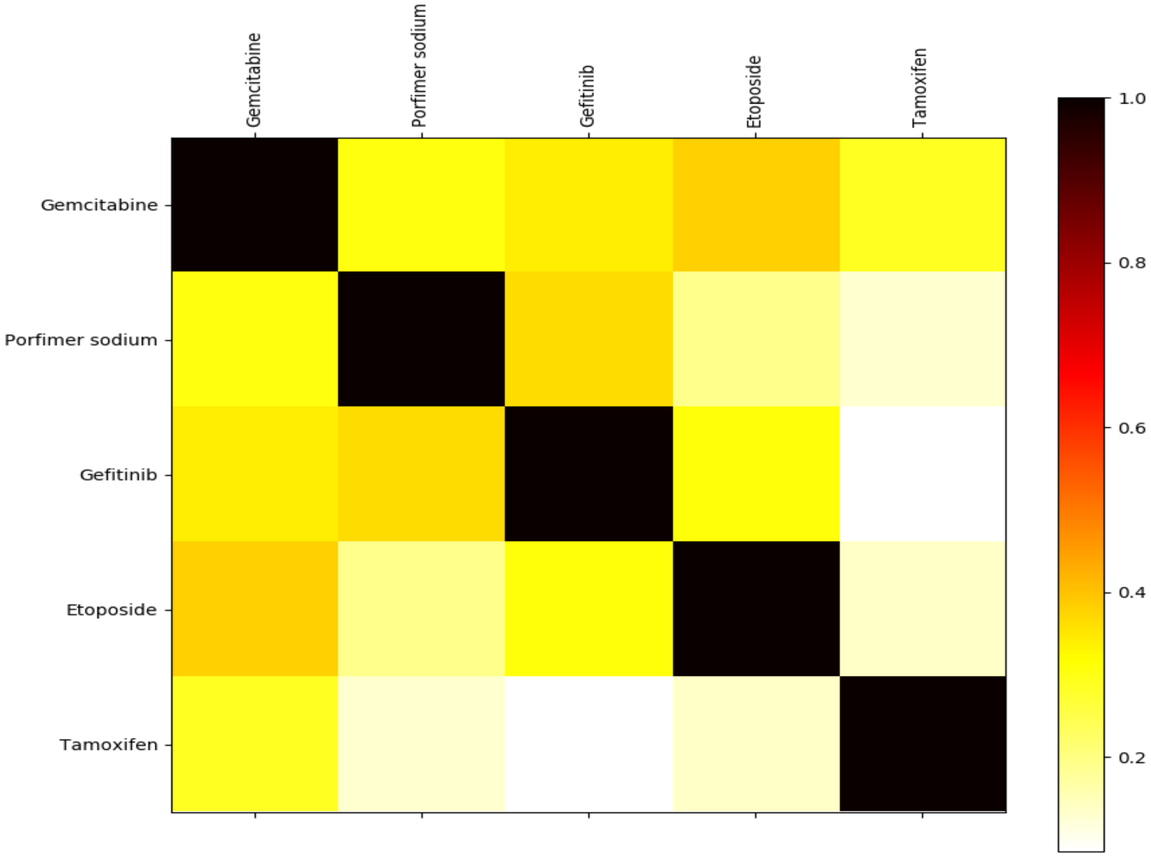

Figure 6.2: Exact Match Example 1

After validating our scoring algorithm to the 5 drugs, we expanded our database by adding 5 more drug records and applied our algorithm on this incrementally expanded dataset. The results are shown in Figure 6.3. We do not see any noticeable difference from the previous experiment. So, We kept incrementally expanding the size of the database until all 50 drug records were included in the similarity assessment. The detailed results are shown in the Appendix D.

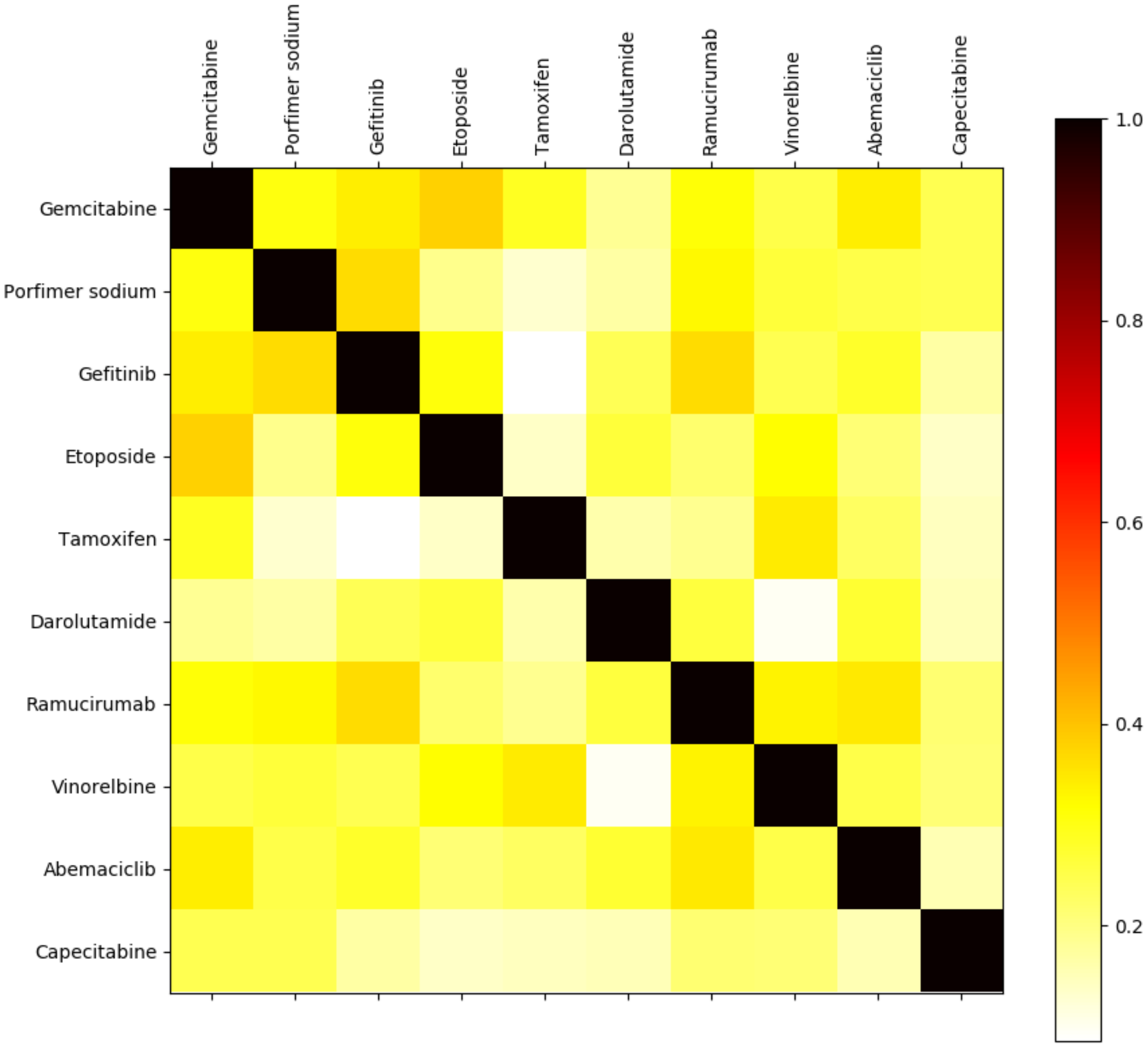

Figure 6.3: Exact Match Example 2

6.4.2 Experimental Results - Best Match Mode

In the Best Match mode, we use one of the drugs “Gemcitabine” in the drug database as a query drug that an analyst uses to calculate the similarity between the query drug and the rest of the drugs in the database. For this experiment we use 10 records, the same 5 drugs in Table 6.3 plus the extra 5 drugs we used in Figure 6.3. We keep the same set of weights for the attributes. In the Best Match (BM) mode, the search options and the semantics for query attributes are specified. The query structure with mode and semantics is shown in Table 6.4.

Table 6.4: Best Match Query

Query Drug	Gemcitabine
Weight	[4, 2, 2, 1, 2]
Mode	[E, B, B, B, E]
Semantics	[x, MB, MB, MB, x]

Table 6.5: Best Match Query Result

Etoposide	0.604
Porfimer sodium	0.559
Ramucirumab	0.552
Tamoxifen	0.515
Vinorelbine	0.492
Abemaciclib	0.478
Gefitinib	0.416
Darolutamide	0.376
Capecitabine	0.344

Scoring functions appropriate to the query mode and semantics are selected from Table 6.2. After computing the similarity measures of all 9 drugs, we choose to include the query drug in the overall ranking. The ranked list of all the 10 drugs are shown in Table 6.5. To better compare the results, we use the same visualization method for the table and it is shown in Figure 6.3. We observe from the ranked list, drug ’Etoposide’ still holds the highest similarity to the drug “Gemcitabine”, whereas the similarity to ’Abemaciclib’ and ’Gefitnib’ are not as significant as before. The reason is the user specified semantics tilts the balance. Due to the preferences enforced by *weights*, ’Ramucirumab’, ’Porfimer sodium’ popped up higher in ranking, as they become more similar to ’Gemcitabine’. This result

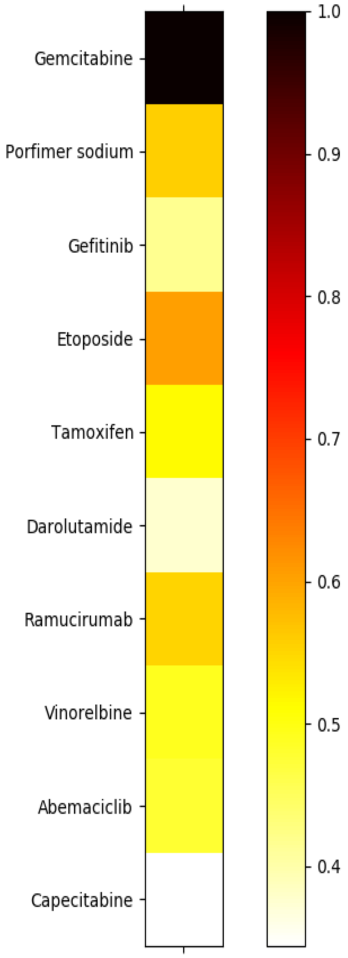

Figure 6.4: Best Match Example

shows the flexibility of our algorithm in producing different rankings to different semantic requirements of analysts.

Finally, we did the experiment over all the 50 drugs in the database. The detailed comparison and result are included in the Appendix D.

6.5 A Summary of Our Method

As we had remarked earlier, it is hard to compare the results produced by the different algorithms because they are based on different hypotheses, they use different drug models, and use different approaches to define and calculate similarity. We acknowledge that most of the authors of the papers we reviewed in Section 6.1 are from Health Sciences or Biomedical Informatics or Medical Research Labs. They have been motivated to study drug-drug similarity from different perspectives and have direct access to real world datasets on which they can experiment. However, they have all used the simple Jaccard measure or its variation on the coded drug model. Because Jaccard measure is purely set-theoretic, although the coded drug models are all "vector models", inaccuracies arise in their calculations. The approach in this thesis is from the point of sound mathematical and rigorous semantic perspective (both domain and user-centric) to designing the similarity functions. So, the methods proposed in the thesis should be viewed as complementary to the work done by those whose work we reviewed earlier, and can be utilized by them. We believe that our methods have the potential to improve the accuracy and semantic relevance. The method proposed in this chapter has the following merits over others:

(1) Ontology-based domain-level semantics proposed in this thesis will improve the accuracy of semantic comparisons of concept terms.

(2) User-centric semantics and preferences for matching and ranking empower the research analysts to vary their preferences and weighing schemes, as appropriate for their objectives. The variety of results they get by varying the parameters might sharpen the predictive power.

(3) Multi-level (both atomic and sets) heterogeneous set of attribute types may provide a sufficiently complete (richer) drug model. Consequently, the results based on our drug model and semantic-based scoring functions may provide better insight to DDI (Drug-drug Interaction) than methods that just use one attribute (such as protein structure) for DDI analysis.

Chapter 7

Patient-Patient Similarity Calculation

In healthcare domain patient similarity assessment is being done by different stakeholder from different perspectives (Sharafoddimi et al., 2017). As an example, administrators and policy makers assess patient satisfaction using an elaborately defined set of questionnaires, and then classify the respondents into clusters where each cluster has a predominant satisfaction level on a set of health services they receive. All respondents in a cluster are regarded as “similar”. In healthcare research, patient similarity assessment is defined as investigating the similarity of patient’s data in terms of one or more of the predictors “disease symptoms. drugs taken, treatment procedures, clinical pathways, and demographics”. In this chapter, we investigate patient similarity assessment based on the drugs they take for cancer symptoms. We use the results from Chapter 6. Thus, our method uses a semantic-based approach (supported by ontology) to determine patient-patient similarity.

We conduct two sets of experiments based on our approach and comment on the observed results. In the first set, the similarity matrix containing the similarity measures between every pair of patients in the database. From this matrix we suggest that the physicians can find quickly the set of all patients similar to a specific patient in the database. This result might facilitate the physician to gain some insight and understanding into the relative progress of the patients in a cohort. As an example, the physician can compare the side effects of patients who are in a “similarity group”, and/or follow their clinical visits to understand whether their recovery paths are similar. In the second, we extract the list of patients from the database who are similar to an "index query patient" and rank them in decreasing order of similarity measure. By gaining insight from this ranked list, a physician might offer personalized prediction to the index patient. We show the results of the two sets of experiments (EM and BM), first for 50 patient records and then incrementally increasing dataset size and repeating the experiments until we reached 1000 patient records. We asses for each experiment both the efficiency of the program and accuracy of results. It is difficult to manually check the accuracy (closeness implied by similarity) exhaustively. However, following the random Delphi type (manual) validation (Okoli & Pawlowski, 2004) we did convince ourselves of the merit of our semantic-based similarity assessment over other methods.

7.1 Related Work

Many researchers have studied patient similarity from different hypothesis on relations among patient features, and have developed methods to calculate patient-patient similarity. The literature on this subject is huge. Fortunately, we found two recent papers (Parimbelli et al., 2018; Sharafoddimi et al., 2017) wherein a comprehensive summary on previous literature is reported. Below, we use their reports. We have already commented on the paper (Zhang et al., 2014) inj Chapter 6, yet we briefly review it below because our work is based on the same hypothesis but our approach to assess patient-patient similarity is based on ontology-based semantics, not just set-theoretic. We also review one paper (Wang et al., 2017) whose goal is to use patient similarity for diabetics patients, whereas we are focusing on cancer patients. The similarity calculation in all these papers are based on *distance-based metrics and inner products*, *neighbourhood-based methods for clustering*, and *set-theoretic measures*. In Chapter 4 we have listed the functions used by these methods, and explained that these approaches use all attributes at once in their functions, and do not allow calculation of similarity at each attribute level. For multifaceted attribute types, semantics-based measures at attribute levels are more suitable. Although attribute level assessment and using them for weighted average of record level similarity are used in (Wang et al., 2017), their methods do not use ontology-based semantics for diabetics disease type. They only use *hierarchy levels* in calculating attribute level scores. But, the definition of "level" in "rooted digraph" ontology is not defined. Presumably, they consider tree structures for ontology and assume "level of a tree node" to be the number of edges along the unique path between it and the root. In summary, no published paper is using drug-drug similarity and the ontology-based methods to assess patient-patient similarity.

7.1.1 Distance-based and Cosine Function Methods

In the survey papers (Parimbelli et al., 2018; Sharafoddimi et al., 2017) the authors have respectively reviewed 1339 and 782 research papers, and selected respectively 22 and 273 papers for reporting the following statistic.

- Cancer is the most frequently considered condition and the methods used by researchers are divided into *clustering*, *dimensionality reduction*, *similarity*, and a *combination of clustering or similarity metrics and supervised approaches.*

- The neighbourhood-based algorithms and distance-based similarity metrics are two of the most frequently used algorithms. Other similarity measures are also used but often they have poor performance and the complexity of calculation is high.

- The concept of similarity defined in most of the papers are "broad and multifaceted". They refer to clinical reports, diagnostic reports, and treatment patterns and often base similarity on textual reports.

- One method utilized the sum of absolute distances on attribute pairs to find the closest class to a index patient.

- Six studies used many statistical measures.

- Several others used cosine function, cluster-based algorithms, and "associations" such as "hospitalization and discharge data" (duration of stay).

Neighbourhood-based modelling methods are easy to implement but their performance is highly dependent on the chosen similarity metric. The Euclidian distance-based similarity metrics are most popular in published papers. They are not good candidates for similarity assessment for the following reasons:

(1) Euclidian distance metric is defined on "geometric modeling" of objects (Tversky, 1977), and the triangle inequality required for such metric is neither necessary nor compatible with the notion of similarity.

(2) It requires all attributes (dimensions) to be of type "numeric".

(3) As the number of attributes increase and/or when the distances become large, the normalization of distance measure becomes problematic.

Although the cluster-based modelling methods have better scalability, the prediction accuracy is lower and sparse clustering results when patients with rare conditions are to be assessed for similarity.

7.1.2 Set-theoretic Similarity Measure

In Chapter 6 we have reviewed the work (Zhang et al., 2014), in which drug-drug similarity is used as the hypothesis for assessing patient-patient similarity towards producing drug personalization for patients. The authors constructed a heterogeneous graph and encoded the three relationships *drug similarity*, *patient similarity*, and *patient-drug prior associations*. They used and compared *chemical structure* and *drug target information* to assess drug similarity. Each drug was represented by an 881-dimensional binary vector, where 1 means the presence and 0 means the absence of a chemical (PubChem structure). The Jaccard Measure, also known as Tanimoto Coefficient, defined by $TC(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$, was used to calculate the similarity between drugs whose PubChem structures are the binary vectors X and Y. So, this measure is a pure set-theoretic measure, expressed as the ratio of the "number of chemicals common to both drugs" to "the total number of chemicals in both drugs". This measure does not take into account the inherent chemical relationships. They collected from Drug Bank (*"DRUGBANK" online*, 2017) the "target protein set" $P(d)$ for each drug d. They defined the similarity between drugs d_1 and d_2 with respect to their target protein sets $P(d_1)$ and $P(d_2)$ using Smith-Waterman sequence alignment score (Okada et al., 2015; Smith & Watermam, 1981). They used ICD9 "diagnostic codes", the standard codes in hospitals to refer to the treatment procedure, with the above two measures to define patient similarity. In summary, they used medical information such as protein sequences, ICD9 codes, and PubChem chemical structures but used only *structural information* based on set and sequence-theoretic calculations to define patient similarity. As opposed to this approach, we use disease ontology, and ontology of ATC codes for drugs to define a semantic similarity to calculate patient-patient similarity.

7.1.3 Attribute-level Score Calculation for Machine Learning

Machine learning methods are used in the paper (Wang et al., 2017), including logistic regression, random forest and kNN. The authors selected the four types of patient information *age*, *gender*, *lab test items*, and *multiple disease diagnosis*. For *age* attribute they used the ratio $\frac{Min(x,y)}{Max(x,y)}$, where Min and Max are respectively functions that compute the minimum and the maximum of the two ages x and y that are compared in the records of two patients. For *gender*, the scoring function returns 1 if the two patients are of the same sex and returns 0 otherwise. For *lab test items*, they assume the following:

- All patients have the same test results, and the same number of test results.
- All test results produce numeric values.
- All values are normalized to lie in the interval $(0, 1)$.

With the above assumptions. they calculate the "normalized value" d of the Euclidian distance D between two terse vectors. That is, $d = \frac{D-Min}{Max-Min}$, where Min and Max are respectively the minimum and the maximum of all Euclidian distances computed over all test vectors. Finally, they assign $1-d$ as the similarity measure between the two test vectors. For *multiple disease diagnosis* attribute, their assumptions and steps are as follows:

(1) Each patient has multiple diseases.

(2) Disease types of patients and the number of diseases of patients may differ.

(3) For any two diseases x and y, from ICD-10 hierarchy they calculate

$$d(x,y) = 1 - D(x,y), \quad \text{where,} \quad D(x,y) = \Big(\frac{Level(NCA(x,y))}{Max\_Level}\Big)$$

In this formula, $NCA(x,y)$ is the *Nearest Common Ancestor* (root of the smallest subtree) of concepts x and y in the hierarchy, and $Level(n)$ is the level of node n in the hierarchy, and Max_level is the maximum level of the tree.

(4) Letting $X = \{x_1, x_2, \cdots, x_n\}$ and $Y = \{y_1, x_2, \cdots, y_m\}$ denote the set of diseases of two patients P_X and P_Y, $X' = X \setminus Y$ and $Y' = Y \setminus X$, they do the following calculations:

- For a fixed $x_i \in X'$, calculate $d(x_i) = \sum_{y \in Y} D(x_i, y)$. This gives the "sum of distances (as defined in Step 3 above) between x_i and the elements of set Y". Calculate the average $Av_1 = \frac{1}{|Y|} \sum_{x_i \in X'} d(x_i)$.
- For a fixed $y_i \in Y'$, calculate $d(y_i) = \sum_{x \in X} D(x, y_i)$. This gives the "sum of distances (as defined in Step 3 above) between y_i and the elements of set X". Calculate the average $Av_2 = \frac{1}{|X|} \sum_{y_i \in Y'} d(y_i)$.
- Define the "multiple diseases similarity" of patients P_X and P_Y as

$$1 - \frac{1}{|X \cup Y|}(Av_1 + Av_2)$$

Having calculated the scores at each attribute level, they assign weights to each attribute and calculate the similarity between patients as the weighted average of the four scores. The following is a list of our observations on this method.

(1) Gender attribute is not necessary for similarity assessment. Because, we can first select from the database all patients with the same gender and then apply similarity assessment on the selected records. This will provide more clear insight on patient similarity within one gender. However, if it is required to compare a specific result across the genders, a better approach is to conduct similarity on "male" records and then on "female records", and then use statistical correlation to investigate how certain results can be compared.

(2) The function $d(x, y)$ that uses "the level of LCA" is problematic, because on a hierarchy with many levels the similarity of nodes that are closer to the root may all be assigned "closer values" which tend to decrease as the level of ontology increases. Moreover, the definition of expressions Av_1 and Av_2 indicate that the distance to concepts that are common to both sets X and Y from concepts that are in X' and Y' are evaluated twice. That might not reflect the "true" similarity. However, this observation needs a more in-depth analysis.

7.2 Motivation

The patient-patient similarity study grants a patient the potential to be treated with personalized care and treatment. With the help of analysis results, a physician can improve the timeliness and quality of care of the index patient while also lowering the medical expenses of the index patient. In the previous section, we have identified many researchers who use the characteristics of drugs, such as chemical targets, indications, side-effects, gene expression profiles (Zhang et al., 2014), and drug categories for patient-patient similarity analysis. In real-world circumstances, the drugs have multiple ways of treatment, such as injection, tablet or powder for solution. For a certain type of treatment method, different dosage may cause different kinds of effectiveness and side effects on patients. Drugs with different strength of dosage may even have separate ATC codes. Thus drug consumption information itself is of high complexity and may vary greatly from one type of patient to the other. Only with the information of drugs and dosage prescribed to a patient, a physician can precisely offer the best care for a patient. So, we take the dosage of prescription information to be of high credibility in drug-drug similarity analysis. It is reported (Parimbelli et al., 2018) that cancer is "most frequently considered condition" for treatment in hospitals. The most studied cancer type for patient similarity is breast cancer. So, we are motivated to propose new similarity functions for assessing similarity of cancer patients type is considered most often by researchers (Parimbelli et al., 2018) and some other reasons mentioned before, we choose to focus on cancer diseases and cancer drugs in this thesis to discuss our methods.

7.3 Similarity Analysis Calculation

We need to consider for each cancer patient the "pseudo patient identifier (PPID)" and set of drugs (SoD) taken by the patient. Hence, the patient record for analysis has only two attributes $PPID$ and SoD. We use $PPID$ only for linking analysis results (and ranked records) to the EHR of the patient in order for physicians to pursue more investigation on patient health. It is not used in similarity analysis. In Chapter 6 we have created a database of 50 drug records. Let us call this database $DRDB$. So, we can choose a random subset of the drug database and assign it to the SoD of a patient. That is, for similarity analysis we have a set SDR of records, where each record in it has two attributes $PPID$ (not used in analysis) and SoD which is a non-empty subset of $DRDB$. The type of SoD is *Set of (DRT)*, where DRT is the "drug record type" defined in Chapter 6. Thus, every patient record P for analysis purpose has a set $D = \{r_1, r_2, \cdots, r_k\}$ of "drug record", $r_i \in DRDB$. For two patients P_i and P_j, let $D_i = \{r_1, r_2, \cdots, r_k\}$ and $D_j = \{s_1, s_2, \cdots, s_l\}$ be the sets of "drug records" in their SoD attribute fields. The similarity $Sim(P_i, P_j)$ of the two patients P_i and P_j is defined using the scoring function 21 for the two sets D_i and D_j. That is,

$$Sim(P_i, P_j) = score(D_i, D_j) = \begin{cases} 1 & D_i = D_j \\ \frac{\sum_{r \in D_i} \sum_{s \in D_j} score(r,s)}{|D_i||D_j|} & D_i \neq D_j \end{cases} \tag{21}$$

Because we have calculated the similarity scores for all pairs of drugs in the database $DRDB$, we look up that table for all scores $score(r, s)$, and calculate $Sim(P_i, P_j)$.

7.3.1 Experimental Results

We generated at random a dataset of 50 patient records. The set of drugs for each patient is a subset of drug records chosen from $DRDB$ database. We conducted two experiments. In the first, we calculated the patient-patient similarity, using the formula in Equation!21 for calculating the similarity between pairs of drug sets. In the second, we ranked the patient records in decreasing order of similarity of patient records for a given patient query record.

7.3.1.1 Patient-Patient Similarity Table Results

We apply Equation21 to the generated 50 patient record dataset to compute a similarity table containing all the possible pairs of similarity between all the patient records. Like we did in Chapter 6, we also use the figure painted with different levels of colour to visualize the similarity values. Results are shown in Figure 7.1.

The first thing we notice is that patient record 11 ('Darolutamide, 600.0') and patient record 16 ('Darolutamide, 600.0') are identical, thus the similarity between them is 1. Patient record 3 ('Cetuximab, 6.0') and patient record 24 ('Cetuximab, 8.0') also have a high similarity, revealing their closeness in dosage. We also notice that the patient record 9 ('Pemetrexed, 25.0', 'Lorlatinib, 100.0') and patient record 19 ('Pemetrexed, 25.0'), patient record 17 ('Nintedanib, 400.0', 'Necitumumab, 64.0', 'Ramucirumab, 40.0') and patient

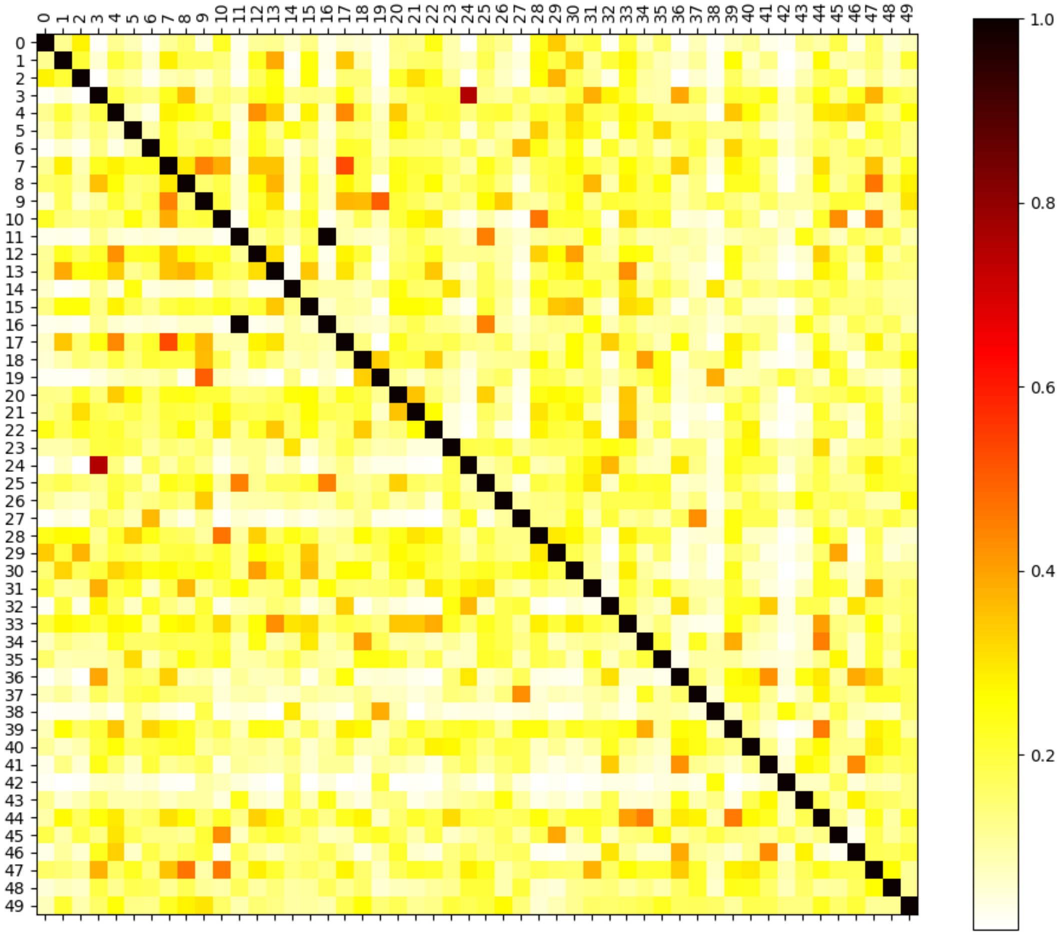

Figure 7.1: Patient-Patient Similarity Example 1

record 32 ('Necitumumab, 64.0') have high similarities since they have one drug with the same dosage in common.

Patient record 27 ('Lenvatinib, 16.0') and patient record 37 ('Lenvatinib, 12.0', 'Cyclophosphamide, 1500.0'), patient record 17 ('Nintedanib, 400.0', 'Necitumumab, 64.0', 'Ramucirumab, 40.0') and patient record 36 ('Ramucirumab, 30.0') have a noticeable similarity because of the same drug they have been prescribed. We also found that patient record 1 ('Necitumumab, 64.0', 'Dacomitinib, 30.0', 'Bevacizumab, 75.0', 'Lorlatinib, 75.0', 'Ceritinib, 600.0') and patient record 32 ('Necitumumab, 64.0') are similar since they contain the same drug with identical dosage. Since the patient record 1 has 5 drugs included, the similarity is not as high as the example we examined before. Across all the high similarity pairs we have examined, we found that patients tend to have clear and distinct similarities when they consume fewer drugs. The patient record 38 ('Topotecan, 0.25') and the patient record 42 ('Cisplatin, 1.0') only contains 1 drug, making it have low similarity across the dataset, but the records similar to them will stand out very clearly.

We also notice that patient record 28 ('Temsirolimus, 75.0', 'Dacomitinib, 15.0', 'Sorafenib, 400.0', 'Nivolumab, 10.0', 'Cabozantinib, 20.0') and patient record 29 ('Exemestane, 25.0', 'Regorafenib, 80.0', 'Temsirolimus, 25.0', 'Ribociclib, 400.0', 'Dactinomycin, 1.5') have acceptable similarity between most of the patients, since they contains 5 drugs, making them easier to meet the patient with same drugs. Though their highest similarity is not as significant as the patient record with fewer drugs, they are more universal.

7.3.1.2 Patient-Patient Query Results

We expanded the patient dataset to 1000 records. We choose the patient record 17 ('Nintedanib, 400.0', 'Necitumumab, 64.0', 'Ramucirumab, 40.0') and patient record 28 ('Temsirolimus, 75.0', 'Dacomitinib, 15.0', 'Sorafenib, 400.0', 'Nivolumab, 10.0', 'Cabozantinib, 20.0') to be our query patients, as patient record 17 contains 3 drugs, and patient record 28 contains 5 drugs. Since the result for one drug query is very easy to predict, queries containing more drugs will give us more informative results on how the algorithm performs on large size real-world datasets.

The top 10 results for using patient record 17 as query patient is shown in Table 7.1. The original query patient record 17 is also included in the table to make it easier to do the comparison. In this result, we can see that of all the top 10 drugs we get, the patient record 36 with only 1 drug and dosage same as the query drug stands out very clearly, and the similarity is higher than other records by a huge amount. In other records, we can notice that the drug-drug similarity records come into play. Patient records 732, 531, and 647 all contain 2 drug consumption record with the 'Necitumumab, 64.0' record is identical to the query drug record, but they have a slight difference in the similarity results. Also, the patient record 472 doesn't have anything identical to the query record, but since the drug 'Nintedanib' (Cancer Name attribute: 'lung non-small cell carcinoma'), 'Necitumumab' (Cancer Name attribute: 'lung non-small cell carcinoma'), 'Ramucirumab'(Cancer Name attribute: 'gastroesophageal junction adenocarcinoma', 'hepatocellular carcinoma', 'colorectal cancer', 'lung non-small cell carcinoma') and 'Ipilimumab'(Cancer Name attribute: 'hepatocellular carcinoma', 'lung non-small cell carcinoma', 'renal cell carcinoma', 'esophageal carcinoma', 'esophageal carcinoma', 'colorectal cancer') all have the same *Cancer Name* attribute for 'lung non-small cell carcinoma', it is ranked high in our dataset. We notice that only the patient record 517 contains 5 drugs prescription, whereas the similarity tends to be distinct when fewer drugs are included.

Table 7.1: Query Results for Patient Record 17

PID	**Prescription**	**Similarity**
17	'Nintedanib, 400.0', 'Necitumumab, 64.0', 'Ramucirumab, 40.0'	1.0
36	'Ramucirumab, 40.0'	0.54781818
732	'Necitumumab, 64.0', 'Avelumab, 80.0'	0.40142424
78	'Cisplatin, 2.0', 'Bevacizumab, 100.0', 'Nintedanib, 100.0'	0.39422475
350	'Necitumumab, 48.0'	0.39422475
531	'Necitumumab, 64.0', 'Bevacizumab, 100.0'	0.38662879
949	'Nintedanib, 100.0'	0.38412121
472	'Ipilimumab, 20.0'	0.35827273
647	'Necitumumab, 64.0', 'Gefitinib, 750.0'	0.34935354
517	'Ramucirumab, 40.0', 'Cabozantinib, 80.0', 'Nintedanib, 100.0', 'Necitumumab, 48.0', 'Pemetrexed, 100.0'	0.34720859
317	'Avelumab, 40.0', 'Necitumumab, 64.0'	0.34165152

We take the query as the drug set in patient record 28 (results are shown in Table 7.2). The first thing we notice is that patient record 134 and 904 have relatively high similarity measure (over 0.40), as all the other drugs have only similarity below or close to 0.35, and patient record 134 and 904 both include an identical record with the query drug. Still, the record with single drug stands out very clearly. And for the least similar drug, we also notice that the records have a single drug included. This suits our findings before. Looking at the top 10 and least 5 similarity records, we notice that all of them have no more than 2 drugs. This means records with more drugs included in it may be in the records that are not included in this category. In general, we can say that when there are more drugs included in records, their similarity tend to be lower.

Our experiments have shown that the drug consumption information is very precise, specific and sensitive to create small differences in patient similarity. The minimal changes in the dosage consumption may affect the overall similarity hugely. In our generated dataset this trend is very noticeable, and it is worth to be tested out in bigger real-world datasets. Also, in our earlier experiment, the drug-drug similarity results are already stored, which

Table 7.2: Query Results for Patient Record 28

PPID	**Prescription**	**Similarity**
28	'Temsirolimus, 25.0', 'Dacomitinib, 30.0', 'Sorafenib, 600.0', 'Nivolumab, 20.0', 'Cabozantinib, 40.0'	1
134	'Nivolumab, 20.0'	0.428931469
904	'Cabozantinib, 40.0'	0.415355711
618	'Nivolumab, 20.0', 'Axitinib, 2.0'	0.355919347
588	'Nivolumab, 20.0', 'Temsirolimus, 75.0'	0.35158345
161	'Temsirolimus, 50.0'	0.336540793
861	'Ipilimumab, 10.0'	0.333282984
667	'Sorafenib, 600.0'	0.329250816
569	'Sorafenib, 400.0'	0.327209557
650	'Dacomitinib, 30.0', 'Atezolizumab, 1680.0'	0.321223105
...	...	...
536	'Cyclophosphamide, 500.0', 'Anastrozole, 2.0'	0.09182987
770	'Darolutamide, 1200.0'	0.090440326
195	'Anastrozole, 2.0', 'Cisplatin, 4.0'	0.089558766
939	'Dactinomycin, 2.0'	0.086165793
886	'Anastrozole, 1.0'	0.054733333

means when we are computing the patient-patient similarity, we only need to access the data inside of the stored data, making it very fast to get the patient-patient similarity results. When we are building our drug dataset, the Drug Bank and other drug databases often only shows the drug information and the drug interaction, which often makes it unclear for researchers like us to determine to what extent, the two drugs are related. It would be very helpful if one of the methods of computing are utilized to compute the similarity beforehand.

Chapter 8

Conclusion and Future Work

The thesis includes four significant contributions to EHR analysis:

- A strictly typed EHR Structure that is both general and extendable.
- Ontology-based and user-centric semantic scoring functions for attributes,
- Drug-drug similarity calculation based on a new drug model.
- Patient-patient similarity calculation, induced by drug similarity.

8.1 EHR Structure

After reviewing the literature on the evolution of EHR and its current structure, we noticed a need to improve and standardize its format. The attribute types and semantics in existing EHRs are not made precise. Moreover, information is scattered around in many records. So, for sharing information for health care delivery and for research many other hospital records other than EHR will be required to be integrated. This process, in the absence of precise semantics of attribute information, is error prone. Motivated by these reasons, we first discussed simple and compound types, as is usually done in formal programming languages software. The strict typing makes the operations on attributes well-defined. Moreover, using higher-order types the construction of record types necessary for EHR modeling becomes formal. Because the operations on higher-order types are well-defined, the correctness of EHR operations is enforced.

The EHR structure is viewed as a "vector" whose components are of different types. A component itself may be a simple attribute (hence simple type) or a compound attribute (hence a compound type). After reviewing the categories of users of EHR and their service-centric requirements we proposed the vector model of EHR that is structured into blocks where within each block one category of information can be included. That is, within each block the attributes are "inter-related" with respect to one category of information. Our EHR model includes attributes necessary to model environmental and social aspects of a patient, as recommended in the WHO patient-centric health model (K. Wan & Alagar, 2015). The above list of blocks may neither be exhaustive nor complete in every detail. Consequently, the choice of attributes, regardless how many and how diverse, may not model a sufficiently complete model of a EHR. To accommodate periodic update of EHR structure it is only necessary to have a flexible implementation, in which the blocks orderings may be maintained while varying the implementations of various blocks depending upon the required efficiency of data search within a block. Currently, based upon a user's access controls, the EHR may be projected to the safe view allowed for the user in order to carry out the services or research analysis. Throughout the thesis, we assume such a EHR model is projected for the analysis of interest.

8.2 Scoring Functions

In comparing two records, assuming that they have the same set of attributes listed in the same order, we compare the pairs of values of each attribute in the record using either EM (exact mode) or BM (best mode) semantics. The options EM and BM are user-centric options. For the BM option a user can add either LB (less is better) or MB (more is better) constraint. In addition, a user can specify a weight for each attribute that denotes the level of importance (significance) of the attribute. The user-centric semantics specified by the user are taken into account together with domain-centric operations of typed attributes in assigning a score for every attribute-pair comparison. In this manner, we integrate both domain-centric semantics and user-centric semantics in the definition of scoring functions. We are motivated to adapt and extend the approach proposed successfully in (Alagar et al., 2018; Alsaig et al., 2017). In their original proposal the scoring functions for user-centric semantics was defined only for “numeric types”. We have extended this to scoring functions for attributes of set types, and for attributes supported by ontology-based semantics.

In Chapter 4 we have pointed out that many researchers defined “semantic distance” on an ontology (a directed graph or a tree), rather than on vector models of objects. Tversky (Tversky, 1977) pointed out that a “distance function” is necessarily based on “vector model of objects” and “similarity” can be based on “feature sets”, and they have different properties. He observed that the distance (metric) function is necessarily “symmetric”, while “similarity” can be “asymmetric”. In an ontology of “concept terms”, two concepts are either related by $Is-A$ semantics or they are not related. That is, an ontology terms form a “partially ordered set” for which symmetric property does not hold. However, distance function must be symmetric. So, there are two ambiguities in their approach. One is defining a distance on “partially ordered set” (rather than on a vector model). The other is, they have ignored “the direction in the ontology graph” in defining “distance between two graph nodes”. With examples, we brought out the inaccuracies in their definitions of “ontology-based distance” measures.

Another issue is, most of researchers (Parimbelli et al., 2018; Sharafoddimi et al., 2017; Vilar et al., 2014; Zhang et al., 2014) have used the set-theoretic Jaccard function $\frac{|X \cap Y)}{|X \cup Y|}$. This function is symmetric. It is a particular instance of Tversky’s most general set-theoretic similarity function (see Chapter 4) that can be tailored either to define a symmetric semantic similarity function or an asymmetric semantic similarity function between two terms of concepts. There are two drawbacks to using the Jaccard function. One is, it captures only “common elements” in the two sets, where “exact match” is used for “equality of elements”. So, many elements that are not “equal”, but they are “semantically related” will be left out. Second drawback is, it cannot be used at attribute level comparison, because “all set elements are used in calculating set union and set intersection”.

To overcome these deficiencies, we have introduced a set of 4 different “ontology-based distance” functions, and one “ontology-based set-theoretic” function to calculate scores between attribute pairs whenever the attribute has an ontology support. We first create a vector model of distance for each element in the ontology, and then use an “inner product style” function to assess the similarity between any two elements in the ontology. So, in essence our distance-based similarity is defined on a vector model that we create. We choose the parameters in the most general function of Tversky to construct an asymmetric similarity function which can be used to calculate the similarity between pairs of terms that are related by partial order. For terms that are not related by partial order, the similarity must be symmetric. So, we can use one the distance functions. Moreover, for LB and MB options the algorithm can choose one of the five similarity functions to produce either the smallest or highest similarity measure.

In summary, the contribution of the thesis to scoring functional are all new.

8.3 Drug-Drug Similarity Calculation

The thesis contributes to a new approach to investigate drug-drug similarity for cancer disease. By restricting to one disease, we are hoping that our method can target drugs similar to a given drug in a more precise manner. The proposed drug model is different from the models used by all researchers, as reviewed in Chapter 6. Most of the previous works focused on one specific element, such as protein sequence or ATC code, to model a drug. The only paper (Ferdousi et al., 2017), where 4 biological elements were used, constructed a binary encoding for each and took the concatenation of these sequences to represent the drug model. They used Jaccard measure as the basis to calculate the drug-drug similarity. We pointed out the inadequacies in this approach. As opposed to these approaches, we used 5 attributes which are inter-related but of different types, and used ontology for some attributes to define scoring functions. The advantage of our approach is that it can be extended to include biological attributes and protein sequences provided proper ontology-based semantic support is provided. Instead of using Jaccard measure, we could calculate semantic-based scoring functions. That might improve the semantic accuracy of drug-drug similarity measure.

8.4 Patient-patient Similarity Calculation

Patient-patient analytic aims to find patients who display similar clinical characteristics to the patients of interest. For personalized medicine, a physician would like to know in advance "whether drug X is likely to be effective for a specific patient Y". We restrict to cancer disease domain, and try to answer this question by using our results on drug-drug similarity for cancer on cancer patients. So, we need to assess the similarity between the characteristics of drugs and patient information. Hence, the patient model we need is the association between a patient record and the set of records of drugs that a patient takes. To make the analysis simple, we assume that each cancer patient has no other disease, and takes a finite number of drugs whose model we have in Chapter 6. We have used the scoring function for set attribute from Chapter 5, and the semantic-based similarity already calculated for drugs in Chapter 6. Because this approach can use the similarity table already constructed for cancer drugs in our database, it is very efficient and can be done for large datasets of patients. Our method can be easily extended for calculating patient similarity when patients have one or more additional disease.

8.5 Future Work

The research approaches in this thesis to structuring EHR, constructing mathematically sound scoring functions, and methods to assess similarity between health records require validation on large datasets maintained in hospitals, research labs, and health care repositories held at governmental organizations. We do not have access to real-life datasets. So, the results of this thesis must be viewed as a contribution to the research community in health care domain. Some specific directions of research are suggested below.

(1) Investigate and implement methods to integrate currently available health care datasets within an organization into the typed vector model proposed in this thesis. Provide access controls for user categories, and facilitate accessing and recovering datasets for specific healthcare delivery and research.

(2) On real-world EHR modeled as above, it is necessary to validate the scoring functions. Experts, using the Delphi model (Grime & Wright, 2004; Okoli & Pawlowski, 2004), should check semantically the similarity table to certify whether or not the scoring functions perform as intended.

(3) Apply the drug-drug similarity method proposed in the thesis by modifying/extending the set of attributes in the drug vector.

(4) Investigate methods that can assess patient-patient similarity based on drug-drug

similarity when patients take drugs for more than one disease. Finding “bottom-up” methods will be a significant direction of work, because of likely superior performance on large datasets. By “bottom-up” we mean, “assessing the similarity (as done in the thesis) for each disease (drug) type and putting together the similarity tables (and ranked lists) for all disease types to assess the overall similarity between patients.”

(5) In a recent work (Weegar & Sundström, 2020), machine learning (ML) approach is used to analyze data that comes from the Karolinska University Hospital in Stockholm from 2007 to 2014. The data is not in the EHR format that we have proposed in this thesis. They have used a data-driven bottom-up approach using natural language analysis with ML to extract features and predict the diagnosis of cervical cancer. The five features extracted by them are *free texts* (hand written notes), *diagnosis codes*, *drug codes*, *lab results* and *procedure codes*. Except the *free text* feature, all other features can be given ontology support. Hence, we can incorporate the features *diagnosis codes*, *drug codes*, *lab results* and em procedure codes as typed attributes in our EHR model. These types are the respective ontology that are available in public domain. However, in our approach we are not currently using “free text”, which can be a semi-structured data. So, a future work would be to extend our approach by linking it to a natural language processing system which will extract k-grams so that we can represent the free text feature as sets of k-grams. We can use an approach similar to (Stefanovic, Kurasova, & Strimaitis, 2019) to assess the similarity between the sets of k-grams in two patient (or drug) records. With this extension, our methods will have the same potential as the ML-based method (Weegar & Sundström, 2020) in assessing similarity between drug records or between patient records that include free text. A comparison between our extended approach and the ML-based approach on large real-life datasets is necessary to assess the relative merits between ML approach and the approach proposed in this thesis.

References

Alagar, V., Alsaig, A., & Mohammad, M. (2018). Ranking composite services. In *Information systems architecture and technology vol. 1: Proceedings of 39th international conference on information systems architecture and technology – "isat" 2018 (eds. l. borzemsk et al.)* (p. 100-110). Advances in Intelligent Systems and Computing: Springer-Verlag, Berlin.

Alsaig, A. (2013). *Semantic-based, multi-featured ranking algorithm for services in service-oriented computing, master of computer science thesis* (Unpublished master's thesis). Concordia University, Montreal, Canada, http://spectrum.library.concordia.ca/978006/.

Alsaig, A., Alagar, V., Mohammad, M., & Alhalabi, W. (2017). A user-centric semantic-based algorithm for ranking services: design and analysis. *Service Oriented Computing and Applications*, *11*(1), 101-120.

Babcock, S., Beverley, J., Cowell, L. G., & Smith, B. (2021). The infectious disease ontology in the age of covid-19. *Journal of Biomedical Semantics*, *12*(13), 1–20.

Bennett, J. O., & Briggs, W. L. (2015). *Using and understanding mathematics: A quantitative reasoning approach: Global edition.* Pearson - Addison Wesley.

Birkhead, G. S., Klompas, M., & Shah, N. (2015). Uses of electronic health records for public health surveillance to advance public health. *Annual Review Public Health*, *36*, 345–359.

Chan, L. W. C., Chan, T., Cheng, L. F., & Mak, W. S. (2010). Machine learning of patient similarity: A case study on predicticting survival in cancer patient after locoregional chemotherapy. In *Proc. ieee international conference on bioinformatics and biomedicine workshops* (pp. 467–470).

Chang, F., & Gupta, N. (2015, December). Progress in electronic medical record adoption in canada. *Canadian Family Physician*, *61*, 1076–1084.

Cheng, L., Li, J., Ju, P., Peng, J., & Wang, Y. (2014). Semfunsim: A new method for measuring disease similarity by integrating semantic and gene functional association. *PLOS ONE*, *9*(6), 1–11.

Chicco, D., & Jurman, G. (2020, November). *Uci machine learning repository: Heart failure clinical records dataset.* UCI Machine Learning Repository. Retrieved from `https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records` (Accessed on 2022-02-21)

Chierichetti, F., & Kumar, R. (2015). Lsh-preserving functions and their applications. *Journal of the ACM*, *62*(5), 1–25.

Cordi, V., Lombardi, P., Martelli, M., & Mascardi, V. (2005). An ontology-based similarity between sets of concepts. In *Woa 2005 ptroceedings* (pp. 16–21).

Cross, V. (2006). Tversky's parameterized similarity ratio model: A basis for semantic relatedness. In *Nafips 2006 proceedings: 2006 annual meeting of the north american fuzzy information processing society* (pp. 1538–1542).

Dale, N., & Walker, H. M. (1996). *Abstract data types: Specifications, implementations, and applications.* D.C. Heath and Company, Lexington, MU.

Daoui, A., Gherabi, N., & Marzouk, A. (2017). An enhanced method to compute the similarity between concepts in ontologyt. In *Advances in intelligent systems and computing 640* (p. 95-107). Springer-Verlag, Berlin.

"drugbank" online. (2017). UCI Machine Learning Repository. Retrieved from `https://go.drugbank.com` (Accessed on 2022-07-21)

Eisler, H., & Ekman, G. (1959). A mechanism of subjective similarity. *Acta Psychologica*, *16*, 1–10.

El-Sappagh, Franda, S., & et al., F. A. (2018). Snomed ct standard ontology based on the ontology for general medical science. *BMC Med Inform Decis Mak*, *18*(78), 1–19.

"FDA". (2015). *Druga@fda-fda approved drugs* (Tech. Rep.). U.S. Food & Drug Administration: <https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm?event=browseBy Letter.page & productLetter=N>.

Ferdousi, R., Safdari, R., & Omidi, Y. (2017). Computational prediction of drug-drug interactions based on drugs functional similarities. *Journal of Biomedical Informatics*, *5*(1), 1–21.

Girardi, D., Wartner, S., Halmerbauer, G., Ehrenmüller, M., & Kosorus, H. (2016). Using concept hierarchies to improve calculation of patient similarity. *Journal of Biomedical Informatics*, *63*, 66–73.

Graham, R., Knuth, D., & Patashnik, O. (1994). *Concrete mathematics: A foundation for computer science*. Addison-Wesley Publishing Company.

Grime, M. M., & Wright, G. (2004). Delphi method. *Wiley StatsRef:Statistics Reference Online*, 1–6.

Gunther, T. D., & Terry, N. P. (2005). The emergence of national electronic health record architectures in the united states and australia: Models, costs, questions. *Journal of Medical Internet*, *7*(1), 1–13.

Harispe, S., Sánchez, D., Ranwez, S., Janaqi, S., & Montmain, J. (2014). A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. *Journal of Biomedical Informatics*, 38–53.

Hastings, J., Ceusters, W., Jensen, M., K, M., & Smith, B. (2012). Reprsenting mental functioning: Ontologies for mental health and disease. In *Towards an ontology of mental functioning: Third international conference on biomedical ontology* (pp. 1–5).

Hecht, J. (2019). Fixing a broken record. *"NATURE"*, *573*, 5114–5116.

Hu, Y., Zhou, M., Shi, H., Ju, H., Jiang, Q., & Cheng, L. (2017). Measuring disease similarity and predicting disease-related ncrnas by a novel method. *BMC Medical Genomics*, *10*, 67–74.

Huang, L., Luo, H., Yang, M., Wu, F., & Wang, J. (2021). Drug-drug similarity measure and its applications. *Briefings in Bioinformatics*, *22*(4), 1–20.

Huang, Y., Wang, N., Liu, H., Zhang, H., Fei, X., Wei, L., & Chen, H. (2019). Study on patient similarity measurement based on electronic medical records. In *Proceedings of international medical informatics association (imia), ios press* (pp. 1484–1485).

Huang, Z., Dong, W., Duan, H., & Li, H. (2013). Similarity measure between patient traces for clinical pathway analysis: problem, method, and applications. In *Artificial intelligence in medicine (aime2013)* (pp. 268–272). Lecture Notes in Computer Science, Volume 7885.

Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (2017, November). *Uci machine learning repository: Heart disease dataset.* UCI Machine Learning Repository. Retrieved from `https://archive.ics.uci.edu/ml/datasets/heart+disease` (Accessed on 2022-02-21)

Kruse, C. S., Stein, A., Thomas, H., & Kaur, H. (2018). The use of electronic health records to support population health: A systematic review of the literature. *Journal of Medical Systems*, *42*(11), 214.

Kunimoto1, R., Vogt1, M., & Bajorath1, J. (2016). Maximum common substructure-based tversky index: an asymmetric hybrid similarity measure. *Comput Aided Mol Des.*, *30*, 523–531.

Larsen, R. R., & Hastings, J. (2018). From affective science to psychiatric disorder: Ontology as a semantic bridge. *Frontiers in Psychiatry*, *9*, 1–13.

Lee, J., Maslove, D. M., & Dubin, J. A. (2015). Personalized mortality prediction driven by electronic medical data and a patient similarity metric. *PLos ONE*, *10*, 1–13.

Lee, S., Xu, Y., Martin, E. A., Doktorchick, C., Zhang, Z., & Quan, H. (2017). Unlocking the potential of electronic health records for health research. *IJPDS Standard Issue*, *70*, 54–64.

Likavec, S., Lombardi, I., & Cena, F. (2015). *Tversky's feature-based similarity and beyond* (Tech. Rep.). Dipartimento di Informatica, Università di Torino.

Mabotuwana, T., Lee, M., & Cohen-Solal, E. V. (2013). An ontology-based similarity measure for biomedical data - application to radiology reports. *Journal of Biomedical Informatics*, *46*, 857–868.

Mathur, S., & Dinakarpandian, D. (2012). Finding disease similarity based on implicit semantic similarity. *Journal of Biomedical Informatics*, *45*, 363–371.

McGlynn, K. A., & London, W. T. (2011, May). The global epidemiology of hepatocellular carcinoma: present and future. *Clinics in liver disease*, *15*(2).

Mc Namara, K. e. a. (2019). Cardiovascular disease as a leading cause of death: how are pharmacists getting involved?. *Integrated pharmacy research & practice*, *8*(1-11). doi: 10.2147/IPRP.S133088

Merriam Webster. (1828). *Merriam webster dictionary.* https://www.merriam-webster.com/dictionary/attribute.

Metzler, D., Dumais, S., & Meek, C. (2007). Similarity measures for short segments of text. In *Advances in information retrieval* (p. 16-27). Springer-Verlag, Berlin.

Miriam Seoane Santos, P. J. G.-L. A. S. A. C., Pedro Henriques Abreu. (2017, November). *Hcc survival data set.* UCI Machine Learning Repository. Retrieved from `https://archive.ics.uci.edu/ml/datasets/HCC+Survival` (Accessed on 2022-02-21)

Mo, R., Ye, C., & Whitefield, P. H. (2013). *Some similarity indices with potential meteorological applications* (Tech. Rep.). Lational Laboratory for Coastal and Mountain Meteorology.

Okada, D., Ino, F., & Hagihara, K. (2015). Accelerating the smith-waterman algorithm with interpair pruning and band optimization for the all-pairs comparison of base sequences. *Journal of Medical Systems*, *321*(16), 1–15.

Okoli, C., & Pawlowski, S. D. (2004). The delphi method as a research tool: An example, design considerations and applications. *Information & Management*, *42*(1), 1–19.

Parimbelli, E., Marini, S., Sacchi, L., & Bellazi, R. (2018). Patient similarity for precision medicine: A systematic review. *Journal of Biomedical Informatics*, *83*, 87–96.

Pokharel, S. (2020). *Electronic health record representation for similarity computing, doctor of philosophy thesis* (Unpublished doctoral dissertation). School of Information Technology and Electrical Engineering, The University of Quuensland, Australia.

Rada, R., H. Mili E, B., & Blettner, E. (1989). Development and applications of a metric on semantic nets. *IEEE Trans Syst Man Cybern*, *19*, 17–30.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 382–407.

Saaty, T. L. (1983). Priority setting in complex problems. *IEEE Transactions on Engineering Management*, *EM-30*(3), 140-155.

Schriml, L., E, E. M., & etal;. (2018). Human disease ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res*, *8*(47), 965–972.

Sharafoddimi, A., Dublin, J. A., & Lee, J. (2017). Patient similarity in prediction models based on health data: A scoping review. *JMIR Medical Informatics*, *83*, 87–96.

Sketris, I. S., Metge, C. J., Ross, J. L., & MacCara, M. E. (2004). The use of the world health organization anatomical therapeutic chemical/defined daily dose methodology in canada. *Drug Information Journal*, *38*(1), 15–27.

Smith, T., & Watermam, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*(1), 195–197.

Stefanovic, P., Kurasova, O., & Strimaitis, R. (2019). The n-grams based text similarity detection approach using self-organizing maps asimilarity measures. *Applied Sciences*, *9*, 1–14.

Struckmann, S., Ernst, M., Fischer, S., Mah, M., Fuellen, G., & M´oller, S. (2021). Scoring functions for drug-effect similarity. *Briefings in Bioinforrmatics*, *22*(3), 1–8.

Sun, J., Wang, F., & Edabollahi, S. (2012). Supervised patient similarity measure of heterogeneous patient records. In *Acm sigkdd exploration newsletter (14:1)* (p. 16 - 24).

Tombros, A., & Rijsbergen, C. J. (2004). Query-sensitive similarity measures for information retrieval. *Knowledge and Information Systems*, *6*, 617–642.

Touran, A., Gransberg, D. D., Molenaar, K. R., Ghavamifar, K., Mason, D. J., & Fithian, L. A. (2009). *A guidebook for the evaluation of project delivery methods - appendix f: Procedures for determining the weights of selection factors.* The National academic Press, Washington D.C.

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327–352.

Tversky, A., & Krantz, D. H. (1970). The dimensional representation and the metric structure of similar data. *Journal of Mathematical Psychology*, *7*, 572–597.

van de Klundert, J., Gorissen, P., & Zeemering, S. (2010). Measuring clinical pathway adherence. *Journal of Biomedical Bioinformatics*, *43*(6), 861–872.

Vilar, S., E, U., Lorberbaum, T., Hripcsak, G., Friedman, C., & Tatonetti, N. P. (2014). Similarity-based modeling in large-scale prediction of drug-drug interactions. *NATURE PROTOCOLS*, *9*(9), 2147–2163.

von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior.* Princeton University Press.

Wan, G. H., Wang, Y. D., & Guo, M. Z. (2006). An ontology-based method for similarity calculation of concepts in semantic web. In *Proceedings of the fifth international conference on machine learning and cybernetics, august 2006* (pp. 541–546).

Wan, K., & Alagar, V. (2015). Context-aware, knowledge-intensive, and patient-centric mobile health care model. In *2015 12th international conference on fuzzy systems and* *knowledge discovery (fskd)* (p. 2253-2260). doi: 10.1109/FSKD.2015.7382303

Wang, N., Huang, Y., Liu, H., Fei, X., Wei, L., Zhao, X., & Chen, H. (2017). Measurement and application of patient similarity in personalized predictive modeling based on electropnic medical records. *Journal Biomedical Engineering Online*, *5*, 1–17.

Watzlaf, V. J. M., Rhia, F., Zeng, X., Jarymowycz, C., & Firouzum, P. A. (2004, January). *Standards for the content of electronic health record* (Vol. 1). Retrieved from `https://bok.ahima.org/`

Weegar, R., & Sundström, K. (2020, 08). Using machine learning for predicting cervical cancer from swedish electronic health records by mining hierarchical representations. *PLOS ONE*, *15*, e0237911. doi: 10.1371/journal.pone.0237911

Whetzel, P., & et al;. (2011, 06). Bioportal: Enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic acids research*, *39*, W541-5. doi: 10.1093/nar/gkr469

"WHO". (2015). *Ic-10* (Tech. Rep.). <http://www.who.int/classifications/icd/en/>.

Wu, T. J., & et al;. (2015). Generating a focused view of disease ontology cancer terms for pan-cancer data integration and analysis. *Database: The Journal of Biological Databases and Curation*, *2015*.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. In *Proceedings of the 32nd annual meeting of the association for computational linguistics* (pp. 133–138).

Zhang, P., Wang, F., Hu, J., & Sorrentino, R. (2014). Towards personalized medicine: Leveraging patient similarity and drug similarity analytics. In *Proceedings of amia joint summits transci. sci.* (pp. 132–136).

Appendix A

Results for General Ontology

The similarity functions Dist_1, Dist_2, Dist_3, Dist_4 and TVS_{asym} proposed in Chapter 4 have been implemented. In this section we show the similarity measures calculated by each method on every pair of concept terms of the ontology (Figure 4.3), compare and comment on the behavior of similarity functions. The following five tables respectively show the similarity measures calculated by the functions $Dist_1$, $Dist_2$, $Dist_3$, $Dist_4$ and TVS_{asym}.

Table A.1: $Dist_1$ Table for Figure 3.3

	C_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_0	1	0.5	0.55	0.583	0.375	0.417	0.5	0.333
C_1	0.5	1	0.5	0.5	0.5	0.5	0.5	0.5
C_2	0.55	0.5	1	0.533	0.45	0.467	0.5	0.433
C_3	0.583	0.5	0.533	1	0.417	0.444	0.5	0.389
C_4	0.375	0.5	0.45	0.417	1	0.583	0.5	0.667
C_5	0.417	0.5	0.467	0.444	0.583	1	0.5	0.611
C_6	0.5	0.5	0.5	0.5	0.5	0.5	1	0.5
C_7	0.333	0.5	0.433	0.389	0.667	0.611	0.5	1

Although there are only 8 concepts in the ontology, the similarity results spread over five tables are hard to compare. All of the results are standard and fall into the range $[0, 1]$. So we use “color saturation” method to visualize the similarity differences between the nodes in each one the five figures Figure A.1, Figure A.2, Figure A.3, Figure A.4, and Figure A.5. The interpretation is, if the color is closer to dark red, the similarity is closer to 1, and when the color is closer to white, the similarity is closer to 0. Through manual inspection we notice two prominent characteristics:

Table A.2: $Dist_2$ Table for Figure 3.3

	C_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_0	1	0.567	0.567	0.611	0.333	0.389	0.5	0.278
C_1	0.567	1	0.52	0.533	0.45	0.467	0.5	0.433
C_2	0.567	0.52	1	0.533	0.45	0.467	0.5	0.433
C_3	0.611	0.533	0.533	1	0.417	0.444	0.5	0.389
C_4	0.333	0.45	0.45	0.417	1	0.583	0.5	0.667
C_5	0.389	0.467	0.467	0.444	0.583	1	0.5	0.611
C_6	0.5	0.5	0.5	0.5	0.5	0.5	1	0.5
C_7	0.278	0.433	0.433	0.389	0.667	0.611	0.5	1

Table A.3: $Dist_3$ Table for Figure 3.3

	C_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_0	1	0.567	0.567	0.611	0.333	0.433	0.5	0.3
C_1	0.567	1	0.52	0.533	0.45	0.48	0.5	0.44
C_2	0.567	0.52	1	0.533	0.45	0.48	0.5	0.44
C_3	0.611	0.533	0.533	1	0.417	0.467	0.5	0.4
C_4	0.333	0.45	0.45	0.417	1	0.55	0.5	0.65
C_5	0.433	0.48	0.48	0.467	0.55	1	0.5	0.56
C_6	0.5	0.5	0.5	0.5	0.5	0.5	1	0.5
C_7	0.3	0.44	0.44	0.4	0.65	0.56	0.5	1

(1) The vector models are “stable”, in the sense that there is not much variation between similarity measures of most concept pairs. In particular we observe that the similarity measures of concept pairs of which one is internal and the other is leaf seem to match (oe very close) on all vector models.

(2) The similarity measures of the set-theoretic semantic function TVS_{asym} are in general much lower than the values produced by the functions (Dist) of the vector models. Given that the set-theoretic function of Tversky (Tversky, 1977) is rigorous and has been used more universally in many disciplines, we tend to believe that the suggestion that we followed from (Harispe et al., 2014) should be faulty.

In the following chapters where we study patient-patient similarity and drug-drug similarity we will be using any one of the vector models for similarity calculation.

Table A.4: $Dist_4$ Table for Figure 3.3

	C_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_0	1	0.5	0.55	0.583	0.375	0.45	0.5	0.35
C_1	0.5	1	0.5	0.5	0.5	0.5	0.5	0.5
C_2	0.55	0.5	1	0.533	0.45	0.48	0.5	0.44
C_3	0.583	0.5	0.533	1	0.417	0.467	0.5	0.4
C_4	0.375	0.5	0.45	0.417	1	0.55	0.5	0.65
C_5	0.45	0.5	0.48	0.467	0.55	1	0.5	0.56
C_6	0.5	0.5	0.5	0.5	0.5	0.5	1	0.5
C_7	0.35	0.5	0.44	0.4	0.65	0.56	0.5	1

Table A.5: $TVSasym$ Table for Figure 3.3

	C_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_0	1	0.625	0.625	0.625	0.455	0.25	0.455	0.217
C_1	0.714	1	0.333	0.333	0.769	0.455	0.25	0.4
C_2	0.714	0.333	1	0.333	0.25	0.455	0.25	0.4
C_3	0.714	0.333	0.333	1	0.25	0.455	0.769	0.4
C_4	0.556	0.833	0.25	0.25	1	0.286	0.2	0.25
C_5	0.333	0.556	0.556	0.556	0.286	1	0.714	0.909
C_6	0.556	0.25	0.25	0.833	0.2	0.625	1	0.556
C_7	0.294	0.5	0.5	0.5	0.25	0.938	0.652	1

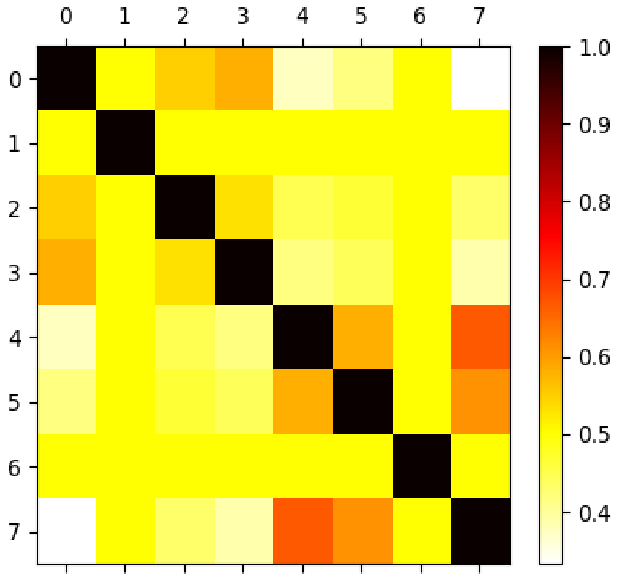

Figure A.1: $Dist_1$ for General Ontology

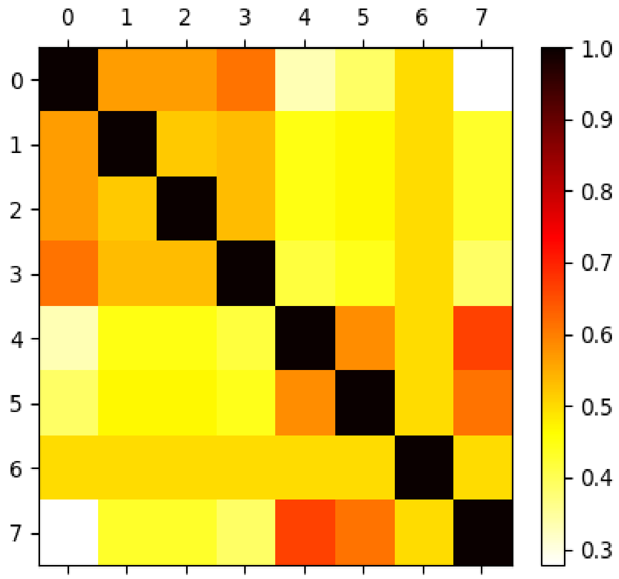

Figure A.2: $Dist_2$ for General Ontology

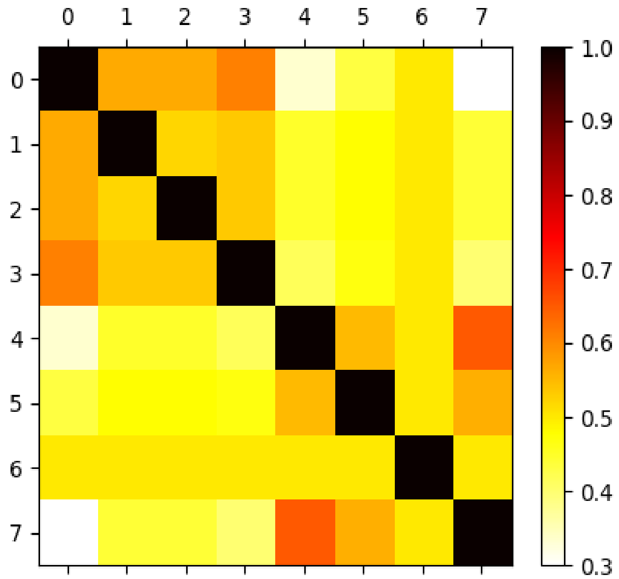

Figure A.3: $Dist_3$ for General Ontology

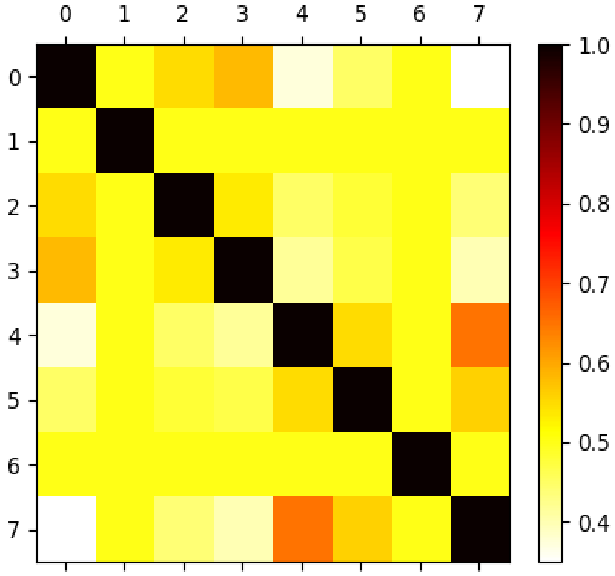

Figure A.4: $Dist_4$ for General Ontology

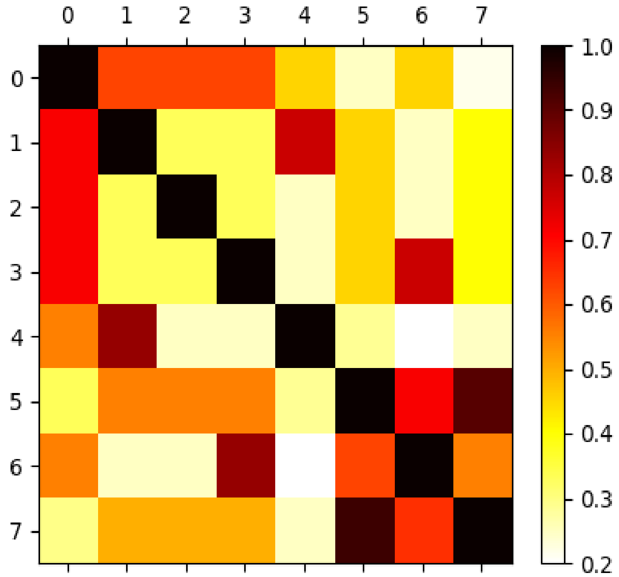

Figure A.5: $TVSasym$ for General Ontology-5

Appendix B

Results for Chapter 4 Case Study

We have implemented the five similarity functions Dist_1, Dist_2, Dist_3, Dist_4 and TVS_{asym} to calculate the similarity measures between every pair of concepts in the emotional ontology case study (Figure 4.7). Since the complete result for each method of this example is a table of size 46x46, displaying it and making an exhaustive manual comparison of it are too hard. So, we skip displaying the five tables, corresponding to the above functions, of similarity measures of pairs of concepts. However, we use the full result for visualization, which are shown in Figure B.1 Figure B.2, Figure B.3, Figure B.4, and Figure B.5. Following our convention, the darker red color means the similarity of this value is closer to 1, and the lighter red color means the similarity of this value is closer to 0, and the shade level indicates gradual decrease in similarity measure. The similarity measure between the same concepts is shown in black color, indicating that the similarity is 1. A visual examination reveals the following properties.

(1) All the four $Dist_n$ functions produce symmetric results. TVS_{asym} function is asymmetric.

(2) The concepts under the same parent have a fairly high similarity in all five tables. Notice the similarity for pairs of the concepts “remembering”, “learning”, “perception”, “disordered thinking”, “fear”, “anger”, “surprise”, and “depressed mood episode” in the visualizations.

(3) For high level of concepts, such as “entity”, all the categorical concepts that it subsumes have high similarity to it. For example, concepts “continuant”, “occurrent”,

“process” and lower level concepts such as “behaviour”, “physiological response to emotion” have this property.

(4) For all functions, the similarity between concept pairs tends to be lower when the level gets lower. When it gets to the specific example of mental disease concepts at the leaf nodes, the similarity is fairly low. But, we find that the similarity is fairly high for the concepts of the same depth for the four distance functions $Dist_n$. So, in the ontology these concepts should be expanded to describe more detailed categories to increase the semantic strength of these concept terms.

(5) The TVS_{asym} function produces high similarity values for pairs of concepts that are directly connected, and produces low similarity values for pairs of concepts that are either not related or far from each other. This means that TVS_{asym} highlights the difference between semantic closeness and semantic separateness.

(6) By comparing the results to the results in Appendix B, we infer that in the emotional ontology (which is a tree) the results using Dist_1 to Dist_4 are close. Also, all the four vector models (the Dist_n functions) tend to reckon the concepts in the same level to have a high similarity. The TVS_{asym} function tends to reckon the concepts in the same path to have a high similarity, thus making the gap between similarity values more noticeable.

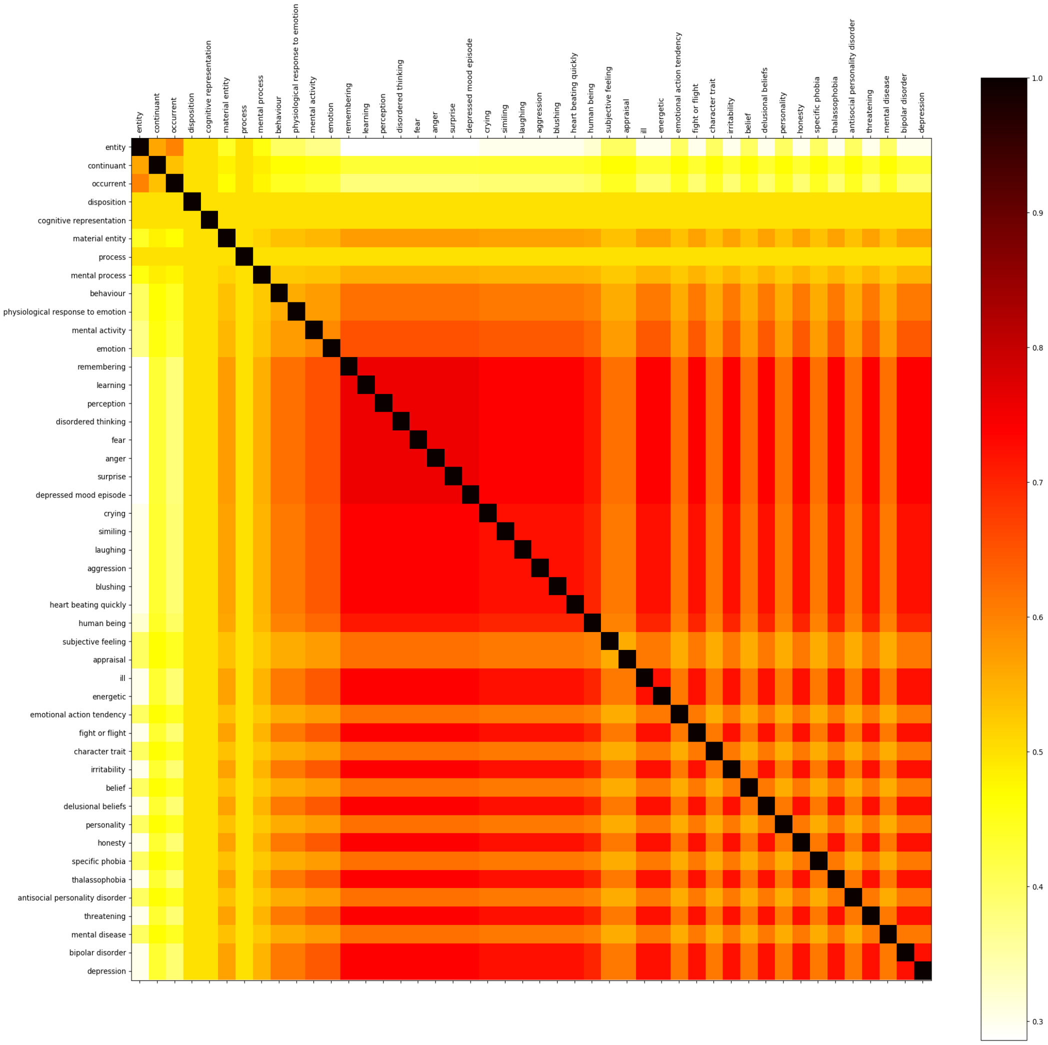

Figure B.1: Dist_1 for Emotional Ontology

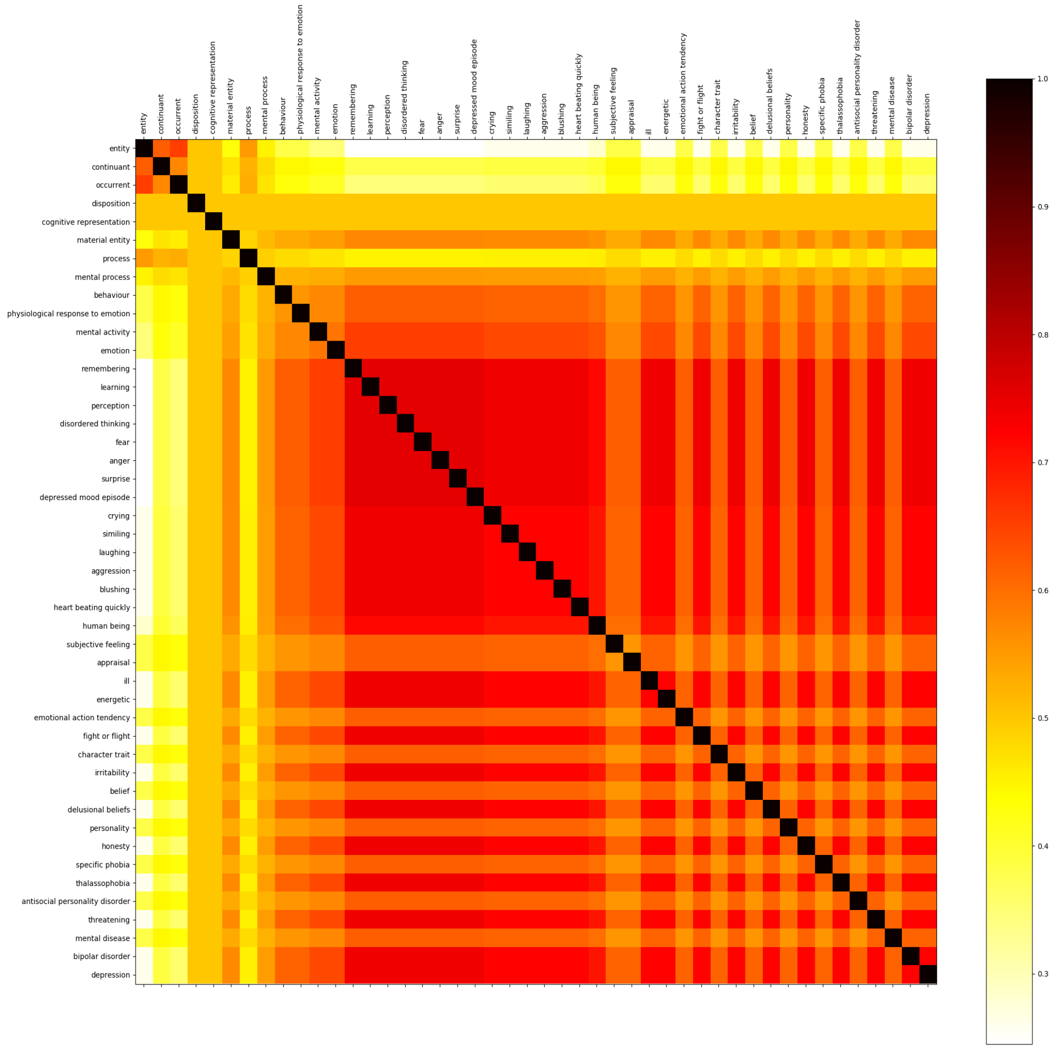

Figure B.2: Dist_2 for Emotional Ontology

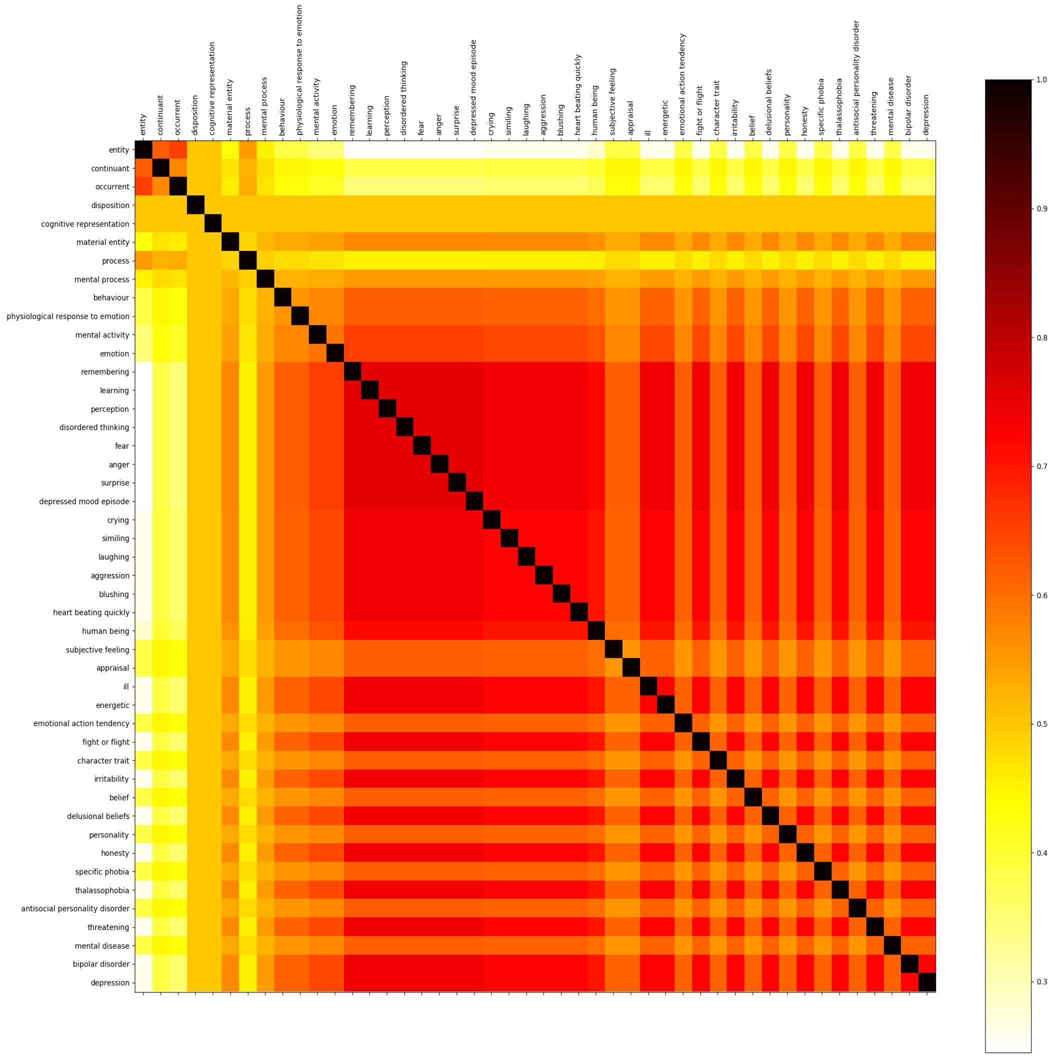

Figure B.3: Dist$_3$ for Emotional Ontology

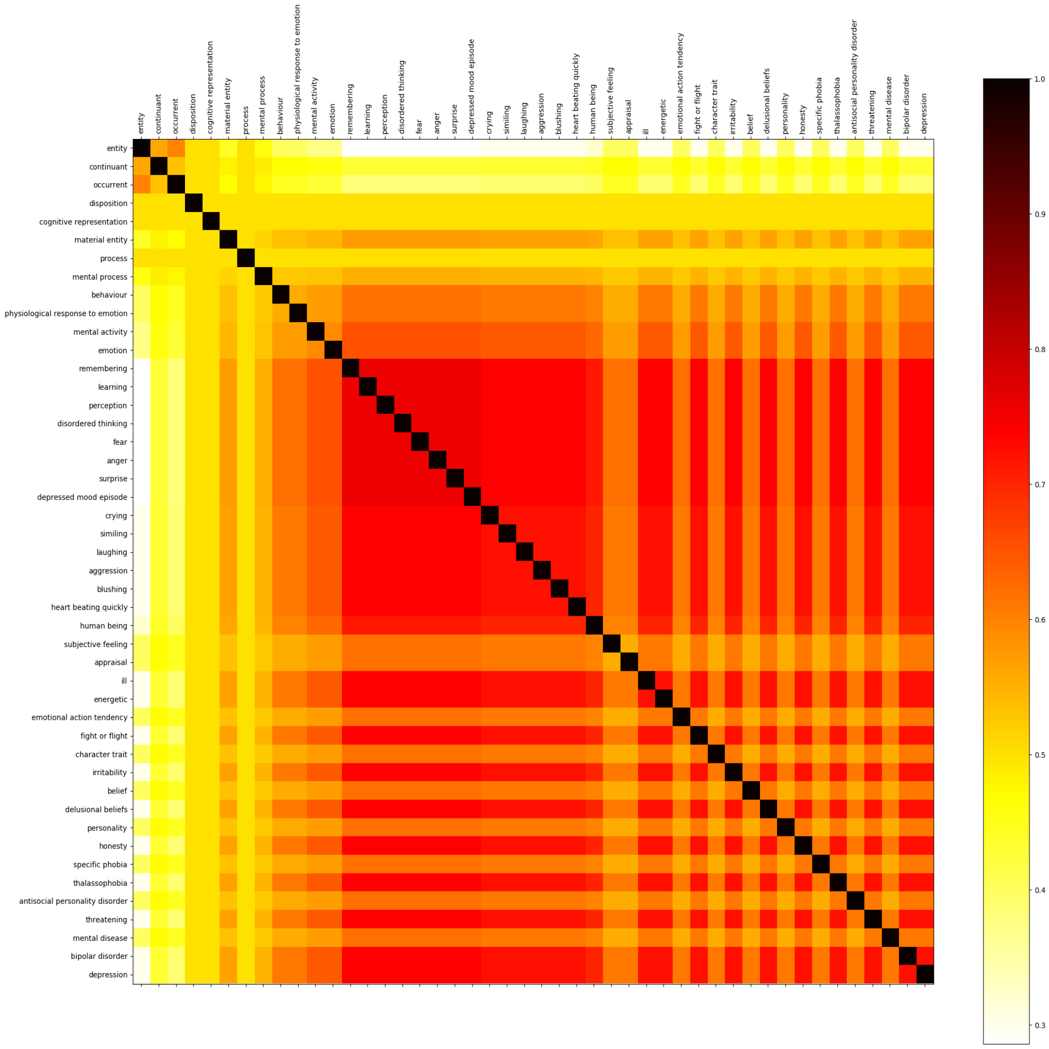

Figure B.4: Dist_4 for Emotional Ontology

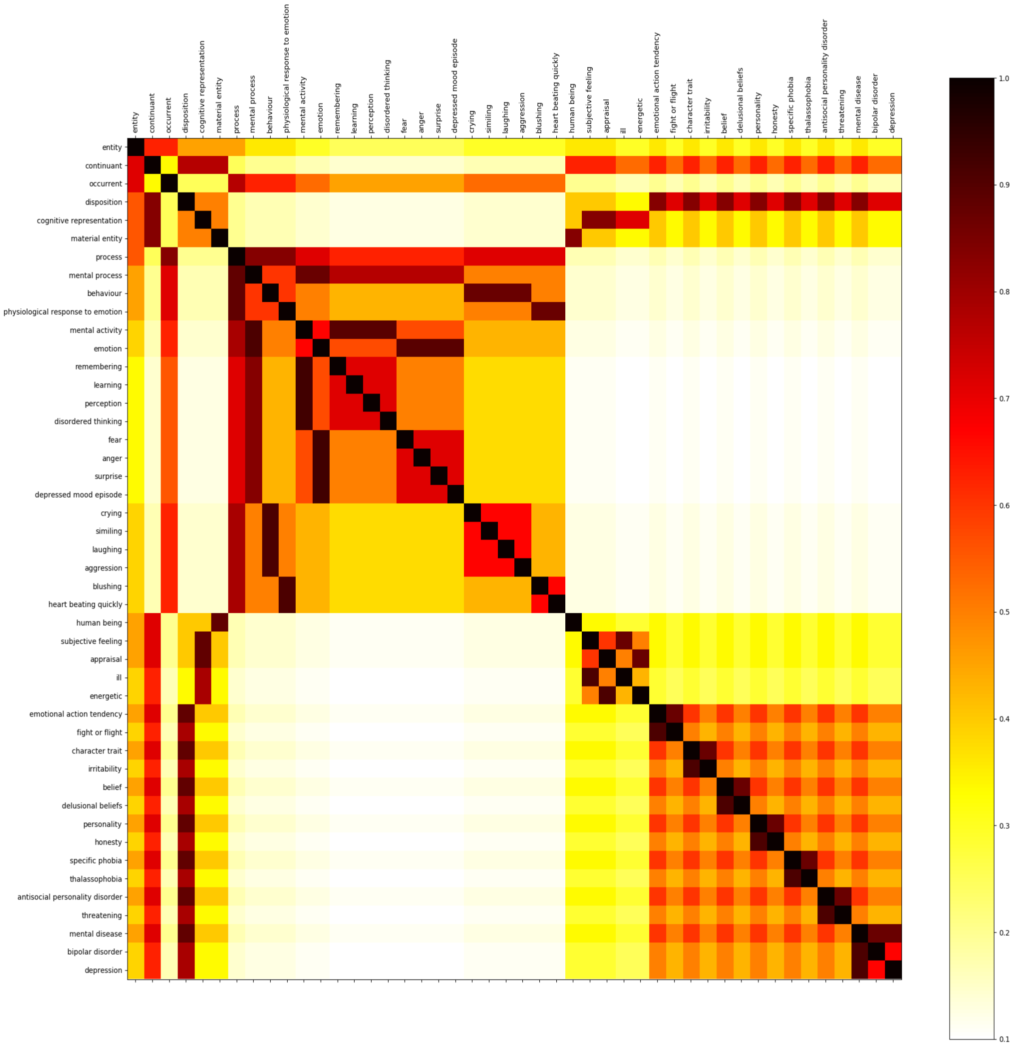

Figure B.5: TVS$_{asym}$ for Emotional Ontology

Appendix C

Results for Sets of Concepts

We have implemented the twenty methods proposed in Chapter 4 for calculating semantic similarity measure for pairs of sets, and for calculating similarity measure for pairs of sets with numerical values. Below, we discuss results on a sample set of examples for Ontology sets.

C.1 Similarity Results for Pairs of Sets from Ontology in Figure 4.3

We create three kinds of set pairs, called <*not related*, *not related*>, <*not related*, *related*> and <*related*, *related*> to compare the similarity functions for pairs of sets. As the name suggests, in the first kind the elements in both sets of the first kind are *not related* by partial order, in the second kind the elements in one of the sets are *not related* by partial order while the elements in the other set are *related* by the partial order, and in the third kind the elements in the both sets are *related* by partial order.

Example 9. *From the Ontology in Figure 4.3 the sets* $S_1 = \{C_2, C_4\}$*, and* $S_2 = \{C_3, C_4\}$ *are chosen. The elements in each set are not related. Table C.1 shows the similarity measures for these two sets, computed by the five functions* $Dist_1, Dist_2, Dist_3, Dist_4, TVS_{asym}$.

Example 10. *From the Ontology in Figure 4.3, the sets* $S_1 = \{C_1, C_2, C_3\}$*, and* $S_2 = \{C_3, C_5, C_7\}$ *are chosen. The elements in set* S_1 *are not related by partial order, and the elements in set* S_2 *are related. Table C.2 shows the similarity measures for these two sets,*

Table C.1: Results for Example 9

	MaxMax	Maxmin	MinMax	MinMin	Average
$Dist_1$	1	0.45	0.533	0.417	0.600
$Dist_2$	1	0.45	0.533	0.417	0.600
$Dist_3$	1	0.45	0.533	0.417	0.600
$Dist_4$	1	0.45	0.533	0.417	0.600
TVS_{asym}	1	0.25	0.333	0.25	0.458

computed by the five functions $Dist_1, Dist_2, Dist_3, Dist_4, TVS_{asym}$.

Table C.2: Results for Example 10

	MaxMax	Maxmin	MinMax	MinMin	Average
$Dist_1$	0.533	0.433	0.533	0.433	0.522
$Dist_2$	0.533	0.5	0.5	0.433	0.529
$Dist_3$	0.533	0.44	0.533	0.44	0.530
$Dist_4$	0.533	0.5	0.5	0.44	0.535
TVS_{asym}	0.455	0.333	0.455	0.333	0.470

Example 11. *From the Ontology in Figure 4.3, the sets* $S_1 = \{C_1, C_7\}$, $S_2 = \{C_3, C_5, C_6\}$ *be two sets are chosen. The elements within each are related, and also a few elements in different sets are related. Table C.3 shows the similarity measures for these two sets, computed by the five functions* $Dist_1, Dist_2, Dist_3, Dist_4, TVS_{asym}$.

Table C.3: Results for Example 11

	MaxMax	Maxmin	MinMax	MinMin	Average
$Dist_1$	0.611	0.467	0.533	0.389	0.500
$Dist_2$	0.611	0.5	0.5	0.389	0.500
$Dist_3$	0.56	0.48	0.533	0.4	0.495
$Dist_4$	0.56	0.5	0.5	0.4	0.493
TVS_{asym}	0.938	0.5	0.455	0.25	0.521

From an inspection of similarity values within each table and across the three tables we observe the following behavior:

- *Table C.1:* We notice that for "MaxMax" choice, the four $Dist_n$ methods produce the same value 1. This is due to the fact that the element C$_4$ exists in both sets, and the maximum of similarity measures returns the value 1. For other methods, all the four $Dist_n$ methods have the same results, because the sets themselves are not related to each other and there's no alternative paths for our calculation. The values for all $Dist_n$ methods are consistently higher than the values produced by TVS_{asym} methods. We believe that this behavior is due to the absence of semantic similarity (lack of partial order) of the set elements. We may conclude that for this case, one of the distance functions $Dist_2, Dist_3, Dist_3$ is a good choice, because the extreme values (like 1 for equal values at element level) does not influence the outcome at set level similarity calculation.

- *Table C.2:* We notice that for "MaxMax" choice, all the four $Dist_n$ methods produce the same values, and these values are quite close to the values of all Dit_n under "MinMax". Similarly, for "MaxMin" and "MinMin" $Dist_n$ methods have values that are close. The values produced by TVS_{asym} are consistently lower than the values produced by $Dist_n$ functions. We infer that this behavior is due to the "unrelatedness" of the elements in the set S_1 and three pairs of elements across the two sets. We observe that due to relateness of pairs in the second set S_2 and many pairs of elements across the two sets, the TVS_{asum} for "Average" method has improved over the previous case.

- *Table C.3:* For both "MaxMin" and "MinMax" choices, all the four $Dist_n$ function values have not changed from the previous case. This means that, increasing relatedness while maintaining distances do not affect the similarity values of these two methods. All $Dist_n$ mathods have "lower" values, compared to the other tables, under "Average" method. In particular, the value for TVS_{asym} function under "Average" has increased, implying that when semantic relateness is strong within a set and across two sets, TVS_{asym} function returns best measures.

C.2 Similarity Results for Pairs of Sets from Emotional Ontology in Figure4.7

In this section we take three types of examples, similar to the types we considered in Section C.1, compute three similarity tables and compare the behavior of the five functions $Dist_1, Dist_2, Dist_3, Dist_4, TVS_{asym}$. Our goal is to investigate whether the behavior we observed in the previous section prevails for the larger Ontology examples.

Example 12. *The concept sets* $S_1 = \{'perception','anger','depressedmoodepisode','blushing'\}$ *and* $S_2 = \{'fightorflight','honesty','threatening','depression'\}$ *are chosen from Figure C.1. In the figure the concepts in the set* S_1 *are shown in "blue", and the concepts belonging to set* S_2 *are shown in "red". Concepts within each set are not related by the partial order. The similarity measure between the sets* S_1 *and* S_2*, calculated by the five functions, are shown in Table C.4.*

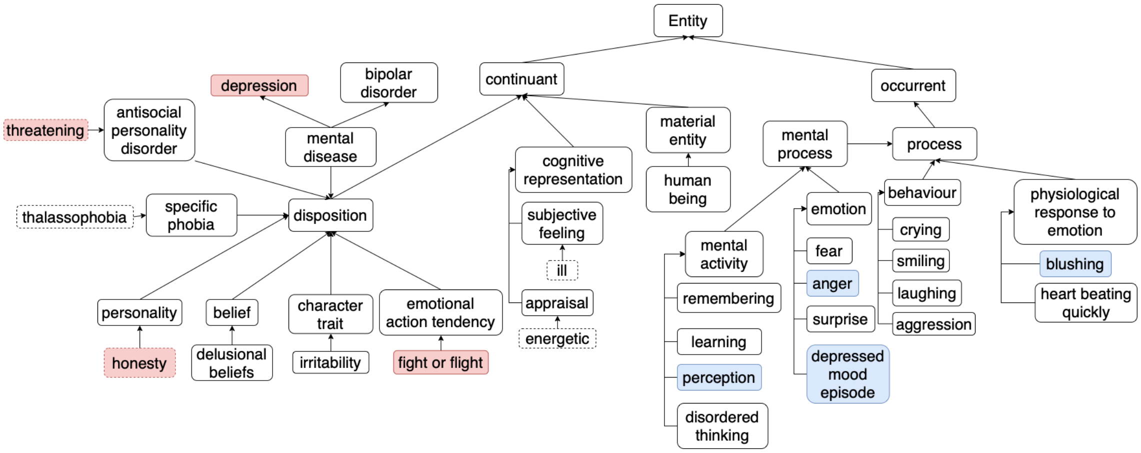

Figure C.1: Ontology for Example 12

Table C.4: Results for Example 12

	MaxMax	Maxmin	MinMax	MinMin	Average
$Dist_1$	0.738	0.738	0.722	0.722	0.733
$Dist_2$	0.738	0.738	0.722	0.722	0.733
$Dist_3$	0.738	0.738	0.722	0.722	0.733
$Dist_4$	0.738	0.738	0.722	0.722	0.733
TVS_{asym}	0.111	0.111	0.1	0.1	0.102

Example 13. *We choose the same set* S_1 *same as in Example 12, in which concepts are not related, and choose* $S_2 = \{'crying','laughing','fear','disorderedthinking'\}$ *in which concepts are related. Figure C.2 is to emphasize the degree of "separateness (distance) and closeness (semantic)" between set elements marked "blue" (set* S_1*), and the set elements marked "red" (set* S_2*). The similarity measure between the sets* S_1 *and* S_2*, calculated by the five functions, are shown in TableC.5.*

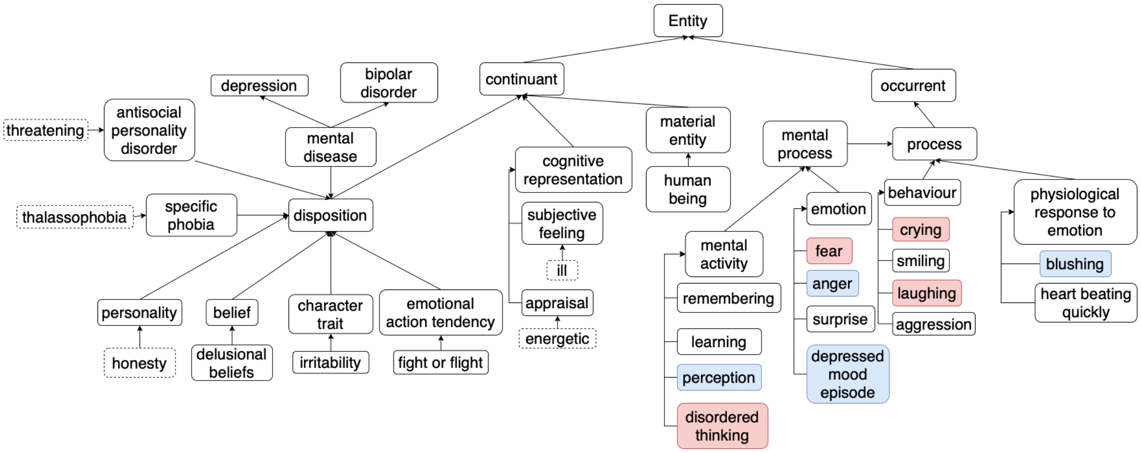

Figure C.2: Ontology for Example 13

Table C.5: Results for Example 13

	MaxMax	Maxmin	MinMax	MinMin	Average
$Dist_1$	0.755	0.738	0.738	0.722	0.742
$Dist_2$	0.755	0.738	0.738	0.722	0.742
$Dist_3$	0.755	0.738	0.738	0.722	0.742
$Dist_4$	0.755	0.738	0.738	0.722	0.742
TVS_{asym}	0.714	0.375	0.729	0.375	0.468

Example 14. *The sets* $S_1 = \{'fear','surprise','blushing','bipolardisorder','delusionalbeliefs',$ $'physiologicalresponsetoemotion','subjectivefeeling','specificphobia'\}$*, and* $S_2 = \{'ill','honesty',$ $'thalassophobia','disposition','crying','heartbeatingquickly','mentaldisease','depression'\}$ *are chosen from FigureC.3. For better comprehension, in the figure we marked in "blue" the concepts belonging to* S_1*, and marked in "red" the concepts belonging to* S_2*. The similarity measure between the sets* S_1 *and* S_2*, calculated by the five functions, are shown in TableC.6*

Table C.6: Results for Example 14

	MaxMax	Maxmin	MinMax	MinMin	Average
$Dist_1$	0.738	0.5	0.738	0.5	0.649
$Dist_2$	0.738	0.5	0.738	0.5	0.649
$Dist_3$	0.738	0.5	0.738	0.5	0.649
$Dist_4$	0.738	0.5	0.738	0.5	0.649
TVS_{asym}	0.375	0.1	0.375	0.1	0.314

From an inspection of values within each table and values across tables, we make the

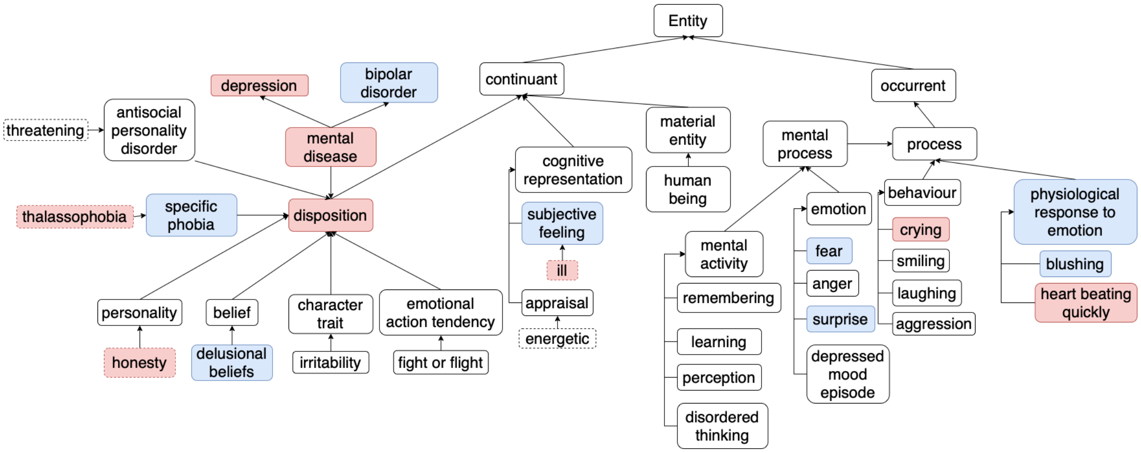

Figure C.3: Ontology for Example 14

following observations to compare the behavior observed in Section C.1.

- From Table C.4 we observe that similarity measures are very high for all $Dist_n$ measures, while TVS_{asym} measure is very low. This is because the concept terms in the sets are "far from related", and they are close to "the leafs". Because of their depth, $Dist_n$ function produce high values, and because of lack of "close relatedness" the semantic measure is low. This observation is consistent with the observations made in Section C.1.

- From Table C.4, and Table C.5 we observe that all the four $Dist_n$ functions produce high values, and these are almost close to each other. Their relative variation is low. However, compared to $theDist_n$ values in them, the $Dist_n$ values in Table C.6 for "Maxmin" and "MinMin" are much smaller. This behavior can be attributed to the closeness of the depth of concepts in the Ontology. Also, we observe that for these cases TVS_{asym} values are high, reflecting their semantic closeness as well.

- The behavior of similarity functions are governed by the "level" (depth) of terms, "the degree of separation" in the graph, and "the extent of semantic binding" suggested by the "the number of nodes subsuming the concept". Such behavior fulfills the "expected behavior" of a similarity function for ontology concepts.

Appendix D

Detailed Results for Drug-Drug Similarity

We explain the *ATC Code* ontology and *Cancer Name* ontology that we use in the experiments. We give the full 50 drug database that we use in our experiments, give the results of the experiments and comment on the similarity results.

Ontology for ATC Codes and Cancer Types

Full ATC ontology for cancer drugs, taken from Drug BankBank (*“DRUGBANK” online*, 2017; "FDA", 2015), has 76 nodes and 75 relations. This ontyology is too big to be shown in the thesis. We use this full ontology for the experiment on the drug database containing 50 drug records. The ATC Ontology, shown in Figure D.1, is extracted from the full ATC ontology in order to just fit the experiment on 10 drug records. This ontology has 26 nodes and 25 relations. For the purpose of the first experiment on 5 drugs, we extracted a sub-ontology of Figure D.1 that has 16 nodes and 17 relations.

Full Cancer ontology, extracted from Bioportal ontology (Whetzel & et al;, 2011), has 138 nodes and 269 relations. This ontology is too big to be shown. We use this ontology for the experiment on 50 drug records. From this ontology we extracted a sub-ontology to fit the experiment on 10 drug records. This ontology has 108 nodes and 207 relations. It is too big to be shown here. From this we extracted the two sub-ontology shown in Figure D.2 and in Figure 6.1. We use them for experiment on 5 drugs.

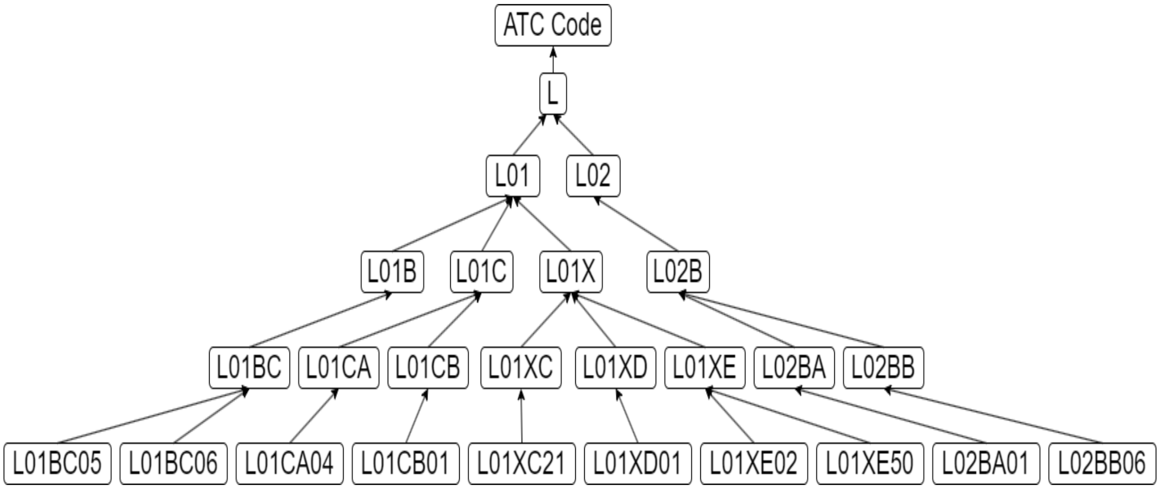

Figure D.1: ATC Code Ontology for 10 Drugs Experiment

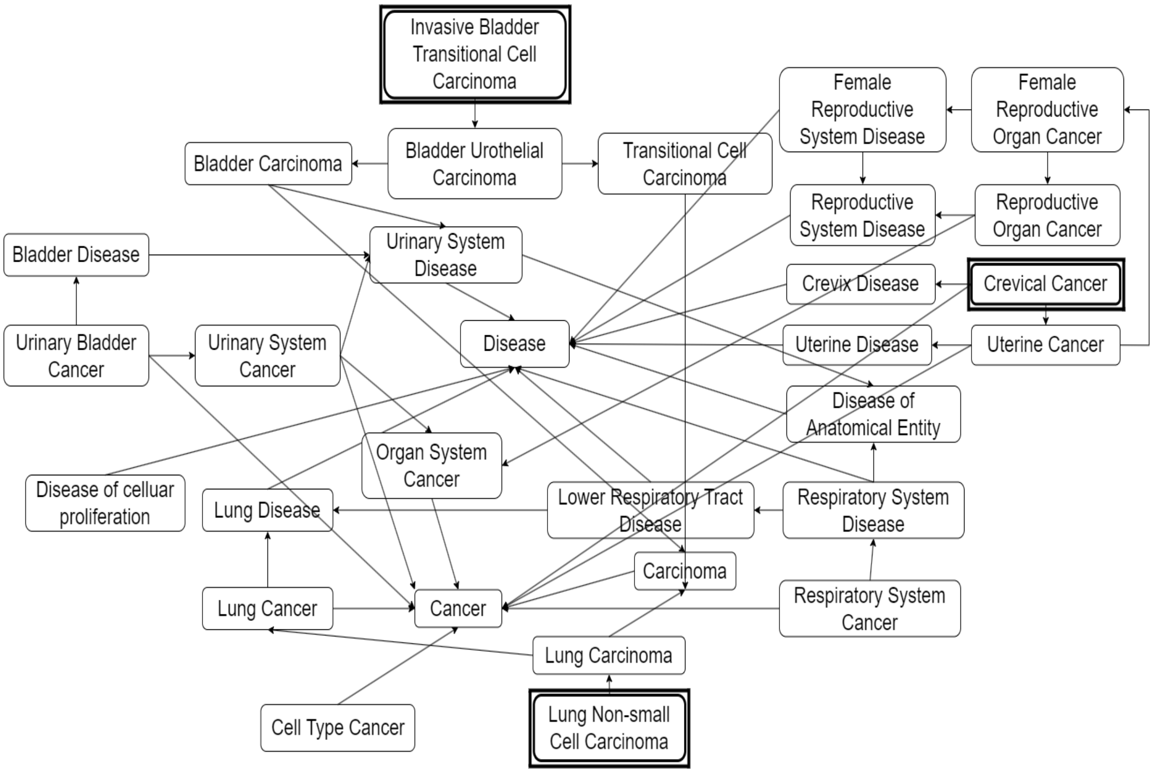
Invasive Bladder Transitional Cell Carcinoma
Female Reproductive System Disease
Female Reproductive Organ Cancer
Bladder Carcinoma
Bladder Urothelial Carcinoma
Transitional Cell Carcinoma
Reproductive System Disease
Reproductive Organ Cancer
Bladder Disease
Urinary System Disease
Crevix Disease
Crevical Cancer
Urinary Bladder Cancer
Urinary System Cancer
Disease
Uterine Disease
Uterine Cancer
Disease of Anatomical Entity
Organ System Cancer
Disease of celluar proliferation
Lung Disease
Lower Respiratory Tract Disease
Respiratory System Disease
Carcinoma
Lung Cancer
Cancer
Respiratory System Cancer
Lung Carcinoma
Cell Type Cancer
Lung Non-small Cell Carcinoma

Figure D.2: Cancer Ontology - for Invasive Bladder Cancer, Cervical Cancer, Lung Cancer

Experimental Results for 50 Drug Records

The results we get from small dataset show reasonable performance of our method in clustering tightly similar drug records. To get more convincing results, we repeated our

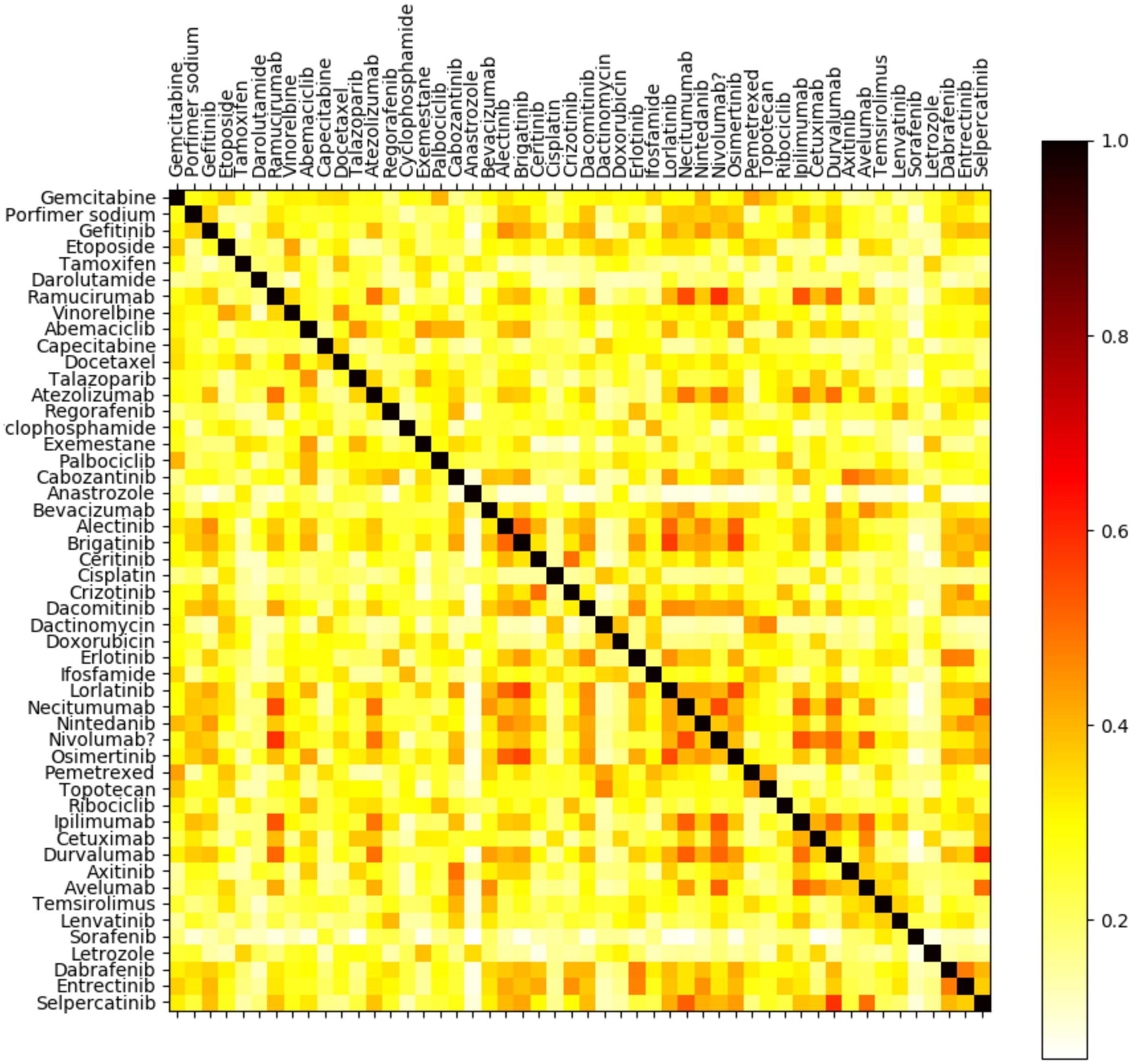

Figure D.3: 50 Drugs Exact Match Example

experiment on all 50 drug records. In Appendix D all the 50 drug records are given.

The difference in similarity measure becomes more noticeable. Tightly similar records retain and improve on the similarity ranking, and loosely similar drugs remain roughly in the same cluster. Figure D.3 shows the clustering for Exact Match applied to all pairs of drugs in the database.

For the “Best Match” experiment we kept the same query that we used for 10 drugs (shown in Table 6.5). We also kept the same set of weights and also used the same drug ’Gemcitabine’ for comparison. The results of the experiment are shown in Figure D.4. In this experiment, the similarity results for the first 10 drugs we used before are roughly the same. We compared the the degree of similarity of these 10 drugs in 10 drug experiment and in the 50 drug experiments. This comparison, shown in Table D.1, reveals that all the similarity values are higher than before, and the order of the ranking stays the same. That is, in some sense the "similarity behaviour" seems "monotonic". This result is surprising and very interesting. Another observation is, the difference between the maximum and the minimum of similarity measures of the 10 drugs in the "10 drug experiment" is 0.24, and in the "50 drugs experiment" is 0.25. That is, the "gap length" that discriminates similarly clustered drugs does not increase much. May be when we repeat the experiments on datasets of increasing sizes, the "strength of similarity" and the "gap length" might *converge*. If that happens, there is justifiable evidence for us to claim that our "similarity calculation" method is *stable* as long as the datasets preserve the "semantics used in our scoring functions."

Table D.1: Best Match Query Result Comparison

Drug Name	**10 Drugs Result**	**50 Drugs Result**
Etoposide	0.552	0.582
Vinorelbine	0.523	0.552
Porfimer sodium	0.499	0.529
Ramucirumab	0.492	0.522
Tamoxifen	0.485	0.502
Abemaciclib	0.429	0.436
Capecitabine	0.416	0.422
Gefitinib	0.356	0.386
Darolutamide	0.312	0.332

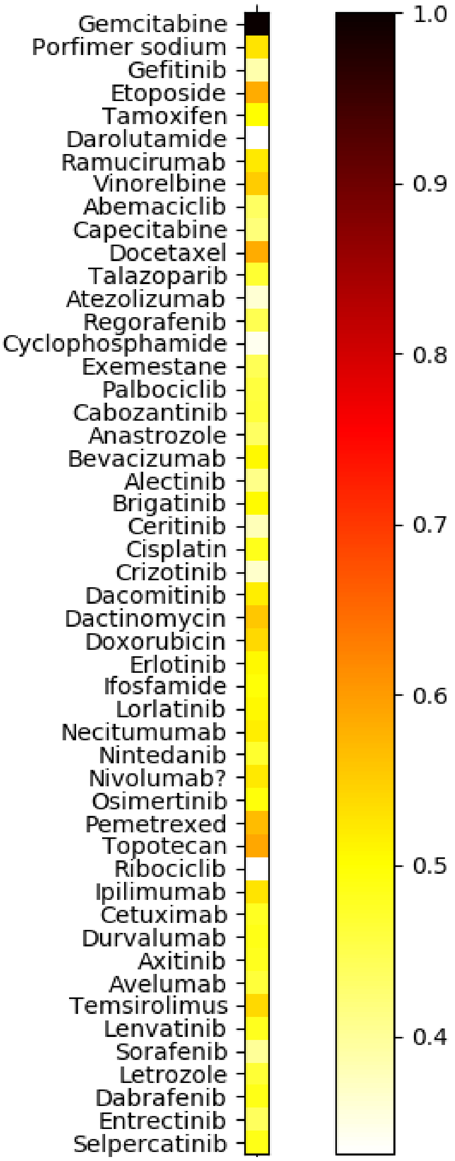

Figure D.4: 50 Drugs Best Match Example

50 Drug Dataset

The drug records in our database data are enumerated below. The vector model of drug records have the attributes <general Name>, <Brand Name>, <ATC Code>, <Cancer Name>, <Dosage>, <Side Effects> in that order. The first 5 drug records are used in the "5 drugs experiment", the first 10 drug records are used in the "10 drugs experiment", and all the 50 drug records are used for the "50 drugs experiment".

<Gemcitabine>,<"Gemzar,Infugem">,<L01BC05>,<"lung non-small cell carcinoma,invasive bladder transitional cell carcinoma,cervical cancer,head and neck carcinoma,lung small cell carcinoma,breast cancer">,<100>,<"increased bleeding,increased infection,increased Thrombosis,increased neutropenic activities,increased immunosuppressive activities,increased myelosuppressive activities">

<Porfimer sodium>,<Photofrin>,<L01XD01>,<"esophageal cancer,lung non-small cell carcinoma">

,<2.5>,<"increased Thrombosis,increased photosensitizing activities">

<Gefitinib>,<Iressa>,<L01XE02>,<lung non-small cell carcinoma,<250,<"increased Thrombosis,increased anticoagulant activities">

<Etoposide>,<"Etopophos,Toposar,Vepesid">,<L01CB01>,<"Merkel cell carcinoma,lung non-small cell carcinoma,ovarian cancer,prostate cancer,retinoblastoma,thymic carcinoma, testis refractory cancer">,<20>,<"increased bleeding,increased Thrombosis,increased infection,

increased cardiotoxicity,increased myelosuppression,

increased neutropenia&thrombocytopenia,increased neutropenia,decreased cardiotoxic activities,increased immunosuppressive activities,increased myelosuppressive activities,increased neutropenic activities">

<Tamoxifen>,<"Nolvadex-D,Soltamox">,<L02BA01>,<"breast cancer,estrogen-receptor positive breast cancer,ovarian cancer",<10>,<"increased bleeding,increased Thrombosis,increased QTc prolongation,decreased cardiotoxic activities,increased QTc-prolonging activities,increased hepatotoxic activities">

<Darolutamide>,<Nubeqa>,<L02BB06>,<castration-resistant prostate carcinoma>,<300>, <increased Thrombosis>

<Ramucirumab>,<Cyramza>,<L01XC21>,<"gastroesophageal junction adenocarcinoma, hepatocellular carcinoma,colorectal cancer,lung non-small cell carcinoma">,<10>,<"increased Thrombosis,increased thrombogenic activities">

<Vinorelbine>,<Vinorelbine>,<L01CA04>,<"lung non-small cell carcinoma,breast cancer,cervical cancer">,<10>,

<"increased neurotoxic,increased bleeding,increased Thrombosis,increased infection,increased cardiotoxic,increased peripheral neuropathy,increased bronchospasm&shortness of

breath&dyspnea,increased myelosuppression,increased neutropenia&thrombocytopenia,increased neurotoxic activities,decreased cardiotoxic activities,increased immunosuppressive activities">
<Abemaciclib>,<Verzenio>,<L01XE50>,<Her2-receptor positive breast cancer>,<50>
,<increased Thrombosis>
<Capecitabine>,<"Ecansya,Xeloda">,<L01BC06>,<"colon cancer,esophageal cancer, hepatobiliary system cancer,colorectal carcinoma,pancreatic cancer,fallopian tube cancer, pancreatic endocrine carcinoma,ovarian cancer,peritoneal carcinoma">,<150>,
<"increased cardiotoxic,increased bleeding,increased Thrombosis,increased infection,increased myelosuppression,increased QTc prolongation,increased neutropenia,increased immunosuppressive activities,increased myelosuppressive activities,increased anticoagulant activities, increased neutropenic activities,decreased cardiotoxic activities">
<Docetaxel>,<Taxotere>,<L01CD02>,<"esophageal cancer,lung non-small cell carcinoma, breast cancer,prostate cancer,bladder carcinoma,gastric adenocarcinoma,head and neck squamous cell carcinoma">,<10>,<"increased bleeding,increased myopathy&rhabdomyolysis& myoglobinuria,
increased cardiotoxic,increased neutropenia&thrombocytopenia,increased Thrombosis,increased hepatotoxic activities,decreased cardiotoxic activities,increased immunosuppressive activities,increased myelosuppressive activities,increased neutropenic activities">
<Talazoparib>,<Talzenna>,<L01XX60>,<breast cancer>,<0.25>,<increased Thrombosis>
<Atezolizumab>,<Tecentriq>,<L01XC32>,<"hepatocellular carcinoma,lung non-small cell carcinoma,ureter carcinoma,lung small cell carcinoma,triple-receptor negative breast cancer">,
<840>,
<"increased Thrombosis,increased thrombogenic activities">
<Regorafenib>,<Stivarga>,<L01XE21>,<"hepatocellular carcinoma,colorectal cancer">,<40>,
<"increased bradycardia,increased Thrombosis,increased neutropenia,increased bradycardic activities">
<Cyclophosphamide>,<Procytox>,<L01AA01>,<"ovary adenocarcinoma,breast cancer,lung cancer">,<500>,<"increased pulmonary toxicity,increased myelosuppression,increased neutropenia, increased neutropenia&thrombocytopenia,increased methemoglobinemia,increased infection,increased granulocytopenia,increased cardiotoxicity,increased bleeding,decreased

cardiotoxic activities,increased fluid retaining&vasopressor activities,increased immunosuppressive activities,increased myelosuppressive activities,increased neurotoxic activities">

<Exemestane>,<Aromasin>,<L02BG06>,<breast cancer>,<25>,<increased Thrombosis>

<Palbociclib>,<Ibrance>,<L01XE33>,<breast cancer>,<75>,<"increased Thrombosis,increased neutropenia,increased myelosuppression,increased infection,increased bleeding">

<Cabozantinib>,<"Cabometyx,Cometriq">,<L01XE26>,<"renal cell carcinoma>,hepatocellular carcinoma">,<20>,

<increased Thrombosis>

<Anastrozole>,<Arimidex>,<L02BG03>,<breast cancer>,<1>,<"increased cardiotoxicity,decreased cardiotoxic activities">

<Bevacizumab>,<"Avastin,Mvasi">,<L01XC07>,<"cervical cancer,colorectal cancer,renal cell carcinoma,lung non-small cell carcinoma,ovarian cancer,fallopian tube cancer,lung non-squamous non-small cell carcinoma,peritoneal carcinoma">,

<25>,<"increased Thrombosis,increased jaw osteonecrosis&anti-angiogenesis,increased cardiotoxicity,increased thrombogenic activities,decreased cardiotoxic activities">

<Alectinib>,<"Alecensa,Alecensaro">,<L01XE36>,<lung non-small cell carcinoma>,<150>,

<increased Thrombosis>

<Brigatinib>,<Alunbrig>,<L01XE43>,<lung non-small cell carcinoma>,<30>,<increased Thrombosis>

<Ceritinib>,<Zykadia>,<L01XE28>,<lung non-small cell carcinoma>,<150>,<"increased QTc prolongation,increased bradycardic activities">

<Cisplatin>,<Platinol>,<L01XA01>,<"testicular cancer,ovarian cancer,urinary bladder cancer">,<1>,<"increased Thrombosis,increased peripheral neuropathy,increased ototoxicity&nephrotoxicity,increased neutropenia, increased nephrotoxicity,increased myelosuppression,increased infection,increased bleeding,decreased cardiotoxic activities,increased bradycardic activities,increased immunosuppressive activities,increased myelosuppressive activities,increased nephrotoxic activities,increased neuromuscular blocking activities">

<Crizotinib>,<Xalkori>,<L01XE16>,<lung non-small cell carcinoma>,<200>,<"increased QTc prolongation,increased bradycardic activities,increased QTc-prolonging activities">

<Dacomitinib>,<Vizimpro>,<L01XE47>,<lung non-small cell carcinoma>,<15>,<"increased Thrombosis,increased neutropenia">

<Dactinomycin>,<Cosmegen>,<L01DA01>,<"ovarian cancer,testicular cancer">, <0.5>,<"increased bleeding,increased Thrombosis,increased neutropenia,increased myelosuppression,increased infection,increased immunosuppressive activities,increased myelosuppressive activities,increased neutropenic activities">

<Doxorubicin>,<"Adriamycin,Doxil,Myocet">,<L01DB01>,<"endometrial cancer,urinary bladder cancer,bronchus carcinoma,ovarian carcinoma,stomach carcinoma,breast cancer">,
<2>,
<"increased Thrombosis,increased neutropenia,increased myelosuppression,increased infection,increased bleeding,increased cardiotoxicity,decreased cardiotoxic activities,increased immunosuppressive activities,increased myelosuppressive activities,increased hepatotoxic& myelosuppressive activities">

<Erlotinib>,<Tarceva>,<L01XE03>,<"pancreatic cancer,lung non-small cell carcinoma">,
<25>,
<"increased Thrombosis,increased neutropenia,increased bradycardia,increased QTc-prolonging activities">

<Ifosfamide>,<Ifex>,<L01AA06>,<"urinary bladder cancer,cervical cancer,head and neck carcinoma,ovarian cancer,lung small cell carcinoma,testicular cancer,thymic carcinoma">,<50>,
<"increased Thrombosis,increased myelosuppression,increased methemoglobinemia,increased infection,increased hemorrhagic cystitis,increased cardiotoxicity,increased bleeding,decreased cardiotoxic activities,increased fluid retaining&vasopressor activities,increased immunosuppressive activities,increased myelosuppressive activities,increased neutropenic activities">

<Lorlatinib>,<Lorbrena>,<L01XE44>,<lung non-small cell carcinoma>,<25>,<increased Thrombosis>

<Necitumumab>,<Portrazza>,<L01XC22>,<lung non-small cell carcinoma>,<16>,
<"increased Thrombosis,increased thrombogenic activities">

<Nintedanib>,<"Ofev,Vargatef">,<L01XE31>,<lung non-small cell carcinoma>,<100>,
<"increased Thrombosis,increased bleeding">

<Nivolumab>,<"Opdivo,Opdualag",<L01XC17>,<"esophageal carcinoma,esophageal cancer,head and neck squamous cell carcinoma,kidney cancer,bladder urothelial carcinoma,lung non-small cell carcinoma,stomach carcinoma,gastroesophageal junction adenocarcinoma, hepatocellular carcinoma,colorectal cancer,gastroesophageal adenocarcinoma">,<10>,<"increased

Thrombosis,increased thrombogenic activities">

<Osimertinib>,<Tagrisso>,<L01XE35>,<lung non-small cell carcinoma>,<40>,<increased Thrombosis>

<Pemetrexed>,<"Alimta,Ciambra,Pemfexy">,<L01BA04>,<"cervical cancer,lung non-squamous non-small cell carcinoma,ovarian cancer">,<25>,<"increased Thrombosis,increased neutropenia,increased myelosuppression,increased infection,increased bleeding,increased immunosuppressive activities,increased myelosuppressive activities,increased neutropenic activities">

<Topotecan>,<Hycamtin>,<L01XX17>,<"lung small cell carcinoma,ovarian cancer,cervical cancer">,<0.25>,<"increased Thrombosis,increased neutropenia,increased myelosuppression,increased infection,increased bleeding,increased immunosuppressive activities,increased myelosuppressive activities,increased neutropenic activities">

<Ribociclib>,<"Kisqali 200 Mg Daily Dose Carton,Kisqali Femara Co-pack">,<L01XE42>, <breast cancer>,<200,<"increased Thrombosis,increased QTc prolongation,increased QTc-prolonging activities">

<Ipilimumab>,<Yervoy>,<L01XC11>,<"hepatocellular carcinoma,lung non-small cell carcinoma,renal cell carcinoma,esophageal carcinoma,esophageal carcinoma,colorectal cancer">, <5>,

<"increased Thrombosis,increased thrombogenic activities">

<Cetuximab>,<Erbitux>,<L01XC06>,<"urinary bladder cancer,breast cancer,head and neck squamous cell carcinoma">,<2>,

<"increased Thrombosis,increased thrombogenic activities">

<Durvalumab>,<Imfinzi>,<L01XC28>,<lung non-small cell carcinoma>,<50>,<"increased Thrombosis,increased thrombogenic activities">

<Axitinib>,<Inlyta>,<L01XE17>,<renal cell carcinoma>,<1>,<increased Thrombosis>

<Avelumab>,<Bavencio>,<L01XC31>,<"renal cell carcinoma,bladder urothelial carcinoma, Merkel cell carcinoma">,<20>,<"increased Thrombosis,increased thrombogenic activities">

<Temsirolimus>,<Torisel>,<L01XE09>,<renal cell carcinoma>,<25>,<"increased Thrombosis,increased neutropenia,increased myelosuppression,increased infection,increased bleeding,increased angioedema,increased immunosuppressive activities,increased myelosuppressive activities,increased QTc-prolonging activities,increased neutropenic activities">

<Lenvatinib>,<Lenvima 10>,<L01XE29>,<"renal cell carcinoma,endometrial carcinoma,

hepatocellular carcinoma">,<4>,<"increased Thrombosis,increased QTc prolongation,increased liver damage,increased bradycardia,increased QTc-prolonging activities">

<Sorafenib>,<Nexavar>,<L01XE05>,<"renal cell carcinoma,hepatocellular carcinoma">,

<200>,

<"increased Thrombosis,increased QTc prolongation,increased neutropenia,increased myelosuppression,increased infection,increased death, increased bradycardia,increased bleeding, increased neutropenic activities,increased QTc-prolonging activities,increased myelosuppressive activities, increased anticoagulant activities,increased immunosuppressive activities">

<Letrozole>,<"Femara,Kisqali Femara Co-pack">,<L02BG04>,<"breast cancer,ovarian cancer">,

<2.5>,<"increased Thrombosis,increased myopathy&rhabdomyolysis&myoglobinuria,increased QTc-prolonging activities">

<Dabrafenib>,<Tafinlar>,<L01XE23>,<lung non-small cell carcinoma>,<50>,<"increased Thrombosis,increased QTc prolongation,increased neutropenia,increased bradycardia,increased photosensitizing activities,increased QTc-prolonging activities">

<Entrectinib>,<Rozlytrek>,<L01XE56>,<lung non-small cell carcinoma>,<100>,<"increased Thrombosis,increased QTc prolongation,increased bradycardia,increased QTc-prolonging activities">

<Selpercatinib>,<Retevmo>,<L01XC33>,<"basal cell carcinoma,skin squamous cell carcinoma,lung non-small cell carcinoma">,<50>,<"increased Thrombosis,increased thrombogenic activities">

Appendix E

Patient-Patient Similarity Record

Patient-patient similarity calculation is very fast because we have already computed drug-drug similarity measures. For the 50 patient records shown in Table E.3, the computing time to calculate the patient-patient similarity is less than 0.2 seconds. For answering a target query for 1000 patient records it takes about 0.1 second. Table E.1 and Table E.2 show sample runtime performance of the implementations for different datasets.

From the tables it is clear that drug-drug similarity calculation takes more time as the number of drugs increases. Although we pre-compute just once the similarity table for pairs of concept terms in an ontology, as the number of drug records increases the number of computations necessary to calculate scoring functions for pairs of sets that include disease values in every pair of records will increase. Because of the limited resource environment in which the current implementation is done, the computing time for fetching pre-computed results from tables stored in external devices inevitable increases. That is the main reason for the high cost shown in Table E.1.

Currently there are at most 200-250 drugs approved by FDA. For example, for cancer there are 243 drugs and for diabetes there are 74 drugs. So, when all pre-computed similarity tables for ontology terms can be maintained in the run-time environment the table lookup will be faster and hence drug-drug similarity calculations for large datasets will take much less time.

Table E.1: Drug Performance Analysis

	Drug Record Number	**Average Runtime**
Drug Similarity Table	5	0.973752975
	10	4.055426788
	50	106.1966344
Drug Query 1 (Gefitinib)	50	15.21731742
Drug Query 2 (Tamoxifen)	50	48.76716757

Table E.2: Patient Performance Analysis

	Patient Record Number	**Average Runtime**
Patient Similarity Table	50	0.1411296
Patient Query (Record 28)	1000	0.1105841

Table E.3: Patient Records

PID	**Prescription**
0	’Entrectinib, 100.0’, ’Dactinomycin, 2.0’, ’Gefitinib, 250.0’, ’Palbociclib, 75.0’
1	’Necitumumab, 64.0’, ’Dacomitinib, 30.0’, ’Bevacizumab, 75.0’, ’Lorlatinib, 75.0’, ’Ceritinib, 600.0’
2	’Regorafenib, 80.0’, ’Nintedanib, 200.0’, ’Abemaciclib, 50.0’, ’Anastrozole, 2.0’, ’Entrectinib, 200.0’
3	’Cetuximab, 6.0’
4	’Osimertinib, 120.0’, ’Sorafenib, 200.0’, ’Ipilimumab, 20.0’
5	’Bevacizumab, 50.0’, ’Ramucirumab, 10.0’
6	’Temsirolimus, 100.0’
7	’Lorlatinib, 100.0’, ’Pemetrexed, 75.0’, ’Ramucirumab, 30.0’
8	’Doxorubicin, 8.0’, ’Alectinib, 450.0’, ’Cetuximab, 4.0’
9	’Pemetrexed, 25.0’, ’Lorlatinib, 100.0’
10	’Nivolumab, 10.0’, ’Palbociclib, 225.0’, ’Etoposide, 20.0’, ’Pemetrexed, 75.0’
11	’Darolutamide, 600.0’
12	’Sorafenib, 400.0’, ’Avelumab, 80.0’, ’Durvalumab, 150.0’, ’Gemcitabine, 300.0’
13	’Necitumumab, 48.0’, ’Brigatinib, 60.0’, ’Exemestane, 75.0’, ’Letrozole, 10.0’, ’Alectinib, 600.0’
14	’Atezolizumab, 840.0’
15	’Cyclophosphamide, 500.0’, ’Exemestane, 50.0’, ’Capecitabine, 450.0’, ’Osimertinib, 80.0’, ’Erlotinib, 50.0’
16	’Darolutamide, 600.0’
17	’Nintedanib, 400.0’, ’Necitumumab, 64.0’, ’Ramucirumab, 40.0’
18	’Pemetrexed, 25.0’, ’Anastrozole, 4.0’, ’Sorafenib, 800.0’
19	’Pemetrexed, 25.0’
20	’Osimertinib, 80.0’, ’Regorafenib, 160.0’, ’Ipilimumab, 5.0’
21	’Ribociclib, 600.0’, ’Anastrozole, 2.0’, ’Erlotinib, 100.0’, ’Darolutamide, 1200.0’, ’Ipilimumab, 5.0’
22	’Alectinib, 150.0’, ’Topotecan, 1.0’, ’Exemestane, 75.0’, ’Anastrozole, 3.0’
23	’Atezolizumab, 3360.0’, ’Talazoparib, 0.25’
24	’Cetuximab, 8.0’
25	’Lorlatinib, 50.0’, ’Tamoxifen, 10.0’, ’Darolutamide, 600.0’
26	’Capecitabine, 150.0’, ’Gefitinib, 1000.0’
27	’Lenvatinib, 16.0’
28	’Temsirolimus, 75.0’, ’Dacomitinib, 15.0’, ’Sorafenib, 400.0’, ’Nivolumab, 10.0’, ’Cabozantinib, 20.0’
29	’Exemestane, 25.0’, ’Regorafenib, 80.0’, ’Temsirolimus, 25.0’, ’Ribociclib, 400.0’, ’Dactinomycin, 1.5’
30	’Cabozantinib, 60.0’, ’Erlotinib, 50.0’, ’Necitumumab, 16.0’, ’Avelumab, 80.0’, ’Axitinib, 4.0’
31	’Darolutamide, 300.0’, ’Alectinib, 150.0’, ’Cetuximab, 6.0’
32	’Necitumumab, 64.0’
33	’Durvalumab, 50.0’, ’Brigatinib, 90.0’, ’Atezolizumab, 1680.0’, ’Exemestane, 75.0’, ’Porfimer sodium, 2.5’
34	’Vinorelbine, 20.0’, ’Sorafenib, 600.0’, ’Cyclophosphamide, 500.0’
35	’Docetaxel, 10.0’, ’Brigatinib, 30.0’, ’Doxorubicin, 4.0’
36	’Ramucirumab, 30.0’
37	’Lenvatinib, 12.0’, ’Cyclophosphamide, 1500.0’
38	’Topotecan, 0.25’
39	’Ipilimumab, 20.0’, ’Temsirolimus, 100.0’, ’Tamoxifen, 40.0’, ’Porfimer sodium, 7.5’, ’Vinorelbine, 20.0’
40	’Ipilimumab, 10.0’, ’Topotecan, 1.0’
41	’Dacomitinib, 45.0’
42	’Cisplatin, 1.0’
43	’Gefitinib, 500.0’
44	’Porfimer sodium, 5.0’, ’Atezolizumab, 2520.0’, ’Talazoparib, 0.75’, ’Sorafenib, 600.0’, ’Vinorelbine, 30.0’
45	’Regorafenib, 40.0’, ’Letrozole, 7.5’, ’Etoposide, 20.0’, ’Temsirolimus, 25.0’
46	’Osimertinib, 120.0’
47	’Cetuximab, 4.0’, ’Ramucirumab, 20.0’, ’Palbociclib, 225.0’, ’Tamoxifen, 30.0’
48	’Tamoxifen, 40.0’, ’Axitinib, 2.0’
49	’Ceritinib, 450.0’, ’Doxorubicin, 2.0’, ’Talazoparib, 1.0’