Computing and Informatics, Vol. 37, 2018, 581–613, doi: 10.4149/cai_2018_3_581

PERSONALIZING A CONCEPT SIMILARITY MEASURE IN THE DESCRIPTION LOGIC $\mathcal{ELH}$ WITH PREFERENCE PROFILE

Teeradaj Racharak

School of Information, Computer, and Communication Technology
Sirindhorn International Institute of Technology, Thammasat University, Thailand
&
School of Information Science
Japan Advanced Institute of Science and Technology, Japan
e-mail: r.teeradaj@gmail.com

Boontawee Suntisrivaraporn

School of Information, Computer, and Communication Technology
Sirindhorn International Institute of Technology, Thammasat University, Thailand
&
Business Intelligence and Data Science, Transformation Group
Siam Commercial Bank Co., Ltd., Thailand
e-mail: boontawee.suntisrivaraporn@scb.co.th

Satoshi Tojo

School of Information Science
Japan Advanced Institute of Science and Technology, Japan
e-mail: tojo@jaist.ac.jp

Abstract. Concept similarity measure aims at identifying a degree of commonality of two given concepts and is often regarded as a generalization of the classical reasoning problem of equivalence. That is, any two concepts are equivalent if and only if their similarity degree is one. However, existing measures are often devised based on objective factors, e.g. structural-based measures and interpretation-based

measures. When these measures are employed to characterize similar concepts in an ontology, they may lead to unintuitive results. In this work, we introduce a new notion called concept similarity measure under preference profile with a set of formally defined properties in Description Logics. This new notion may be interpreted as measuring the similarity of two concepts under subjective factors (e.g. the agent's preferences and domain-dependent knowledge). We also develop a measure of the proposed notion and show that our measure satisfies all desirable properties. Two algorithmic procedures are introduced for top-down and bottom-up implementation, respectively, and their computational complexities are intensively studied. Finally, the paper discusses the usefulness of the approach to potential use cases.

Keywords: Concept similarity measure, semantic web ontology, preference profile, description logics

Mathematics Subject Classification 2010: 68-T30

1 INTRODUCTION

Most Description Logics (DLs) [1] are decidable fragments of First Order Logic (FOL) with clearly defined computational properties. DLs are the logical underpinning of the DL flavor of the ontology languages OWL and OWL 2. The advantage of this close connection is that the extensive DLs literature and implementation experiences can be directly exploited by OWL tools. More specifically, DLs provide unambiguous semantics to the modeling constructs available in OWL DL and OWL 2 DL. These semantics make it possible to formalize and design algorithms for a number of reasoning services, which enable the development of ontology applications to become prominent. For instance, ontology classification (or ontology alignment) organizes concepts in an ontology into a subsumption hierarchy and assists in detecting potential errors of a modeling ontology. Though this subsumption hierarchy inevitably benefits ontology modeling, it merely gives two-valued responses, i.e., inferring a concept is subsumed by another concept or not. However, certain pairs of concepts may share commonality even though they are not subsumed. Thus, a considerable amount of research effort has been devoted on measuring similarity of two given concepts, i.e. *concept similarity measure*.

Basically, a concept similarity measure (abbreviated as CSM) is a function mapping from a concept pair to a unit interval (i.e. $0 \leq x \leq 1$ for any real number x). The higher the value is mapped to, the more likely similarity of them may hold. Intuitively, the value 0 can be interpreted as *total dissimilarity* whereas the value 1 can be interpreted as *total similarity* or *equivalence*. Hence, one may regard CSM as a generalization of the classical reasoning problem of equivalence. It plays a major role in the discovery of similar concepts in an ontology. For example, it is employed in bio-medical ontology-based applications to discover functional similarities of gene [2],

it is often used by ontology alignment algorithms [3]. There is currently a significant number of measures in DLs. Prominent examples are [4, 5, 6, 7, 8, 9]. However, these measures are devised based on objective factors. For example, they use the structure (or the interpretation) of concept descriptions to measure. When these measures are employed to characterize similar concepts in an ontology, they may lead to unintuitive results. The following example illustrates that using objective-based measures may not suffice to answer the agent's request.

Example 1. An agent A wants to visit a place for doing some active activities. At that moment, he would like to enjoy walking. Suppose that a place ontology has been modeled as follows:

$$\begin{aligned}
\mathsf{ActivePlace} &\sqsubseteq \mathsf{Place} \sqcap \exists \mathsf{canWalk}.\mathsf{Trekking} \sqcap \exists \mathsf{canSail}.\mathsf{Kayaking}\\
\mathsf{Mangrove} &\sqsubseteq \mathsf{Place} \sqcap \exists \mathsf{canWalk}.\mathsf{Trekking}\\
\mathsf{Beach} &\sqsubseteq \mathsf{Place} \sqcap \exists \mathsf{canSail}.\mathsf{Kayaking}\\
\mathsf{canWalk} &\sqsubseteq \mathsf{canMoveWithLegs}\\
\mathsf{canSail} &\sqsubseteq \mathsf{canTravelWithSails}
\end{aligned}$$

Suppose that a measure used by that Agent A considers merely the objective aspects, it is reasonable to conclude that both Mangrove and Beach are equally similar to the concept ActivePlace. However, by taking into account also the agent's preferences, Mangrove appears more suitable to his perception of ActivePlace at that moment. In other words, he will not be happy if an intelligent system happens to recommend him to go for a Beach.

The example shows that preferences of the agent play a decisive role in the choice of alternatives. In essence, when the choices of an answer are not totally similar to a concept in question, a measure may need to *be tuned* by subjective factors, e.g. the agent's preferences. Another example is shown in [10] on the experiment of the measure sim against SNOMED CT[1], which is one of the largest and the most widely used medical ontologies currently available. It reports that roleGroup and the SNOMED CT top concept SCT-Top can unintentionally increase the degree of similarity. By augmenting that knowledge, the experiment could produce more accurate outputs. The main purpose of this paper is to investigate the use of concept similarity measure under the agent's preferences. As a result, the advantages of our approach are fourfold (cf. Section 4 and Section 5). Firstly, it formalizes the notion of concept similarity measure under the agent's preferences and identifies its desirable properties. Secondly, inspired by the skeptical and credulous measures in [5], when used under different agent's preferences, our theory corresponds to different types of a rational agent, i.e., it has ordering when used by different agents. Thirdly, it presents the similarity measure sim^{π} with mathematical proofs on the satisfaction

[1] `http://bioportal.bioontology.org/ontologies/SNOMEDCT`

of those properties. Lastly, it presents two algorithmic procedures for implementing the measure, viz. a top-down and a bottom-up versions of the proposed measure.

Our developed measure sim^{π} is driven by the structural subsumption characterization by means of tree homomorphism. It is worth to mention that Baader proposes this idea in [11, 12] for $\mathcal{ELH}$ w.r.t. an unfoldable TBox, i.e., the subsumption is characterized by means of an existence of a homomorphism in the reverse direction. The notion of homomorphism degree is originally introduced in [13] and employed at the heart of similarity measure for $\mathcal{EL}$. This idea is extended at the heart of *concept similarity measure under the agent's preferences* for $\mathcal{ELH}$. Preliminary studies of this applicability are reported in our proceeding papers [14, 15]. It should be noted that our measure we introduced, i.e. sim^{π}, may look similar to the measure proposed in [16] in a sense that both are recursive definitions for the same DL $\mathcal{ELH}$; however, they are radically different. These are caused by the distinction of their inspirations and we discuss those points in Section 7. Preliminary, empirical evaluation, and the conclusion are discussed in Section 2, Section 6, and Section 8, respectively.

2 PRELIMINARIES

In this section, we review the basics of the Description Logic $\mathcal{ELH}$ and the problem of concept similarity measure including the measure sim, which is extended to the development of sim^{π} (originally introduced in [15]).

2.1 Description Logic $\mathcal{ELH}$

We assume countably infinite sets CN of concept names and RN of role names that are fixed and disjoint. The set of concept descriptions, or simply concepts, for a specific DL $\mathcal{L}$ is denoted by $\mathsf{Con}(\mathcal{L})$. The set $\mathsf{Con}(\mathcal{ELH})$ of all $\mathcal{ELH}$ concepts can be inductively defined by the following grammar:

$$\mathsf{Con}(\mathcal{ELH}) ::= A \mid \top \mid C \sqcap D \mid \exists r.C$$

where $\top$ denotes the *top concept*, $A \in \mathsf{CN}$, $r \in \mathsf{RN}$, and $C, D \in \mathsf{Con}(\mathcal{ELH})$. Conventionally, concept names are denoted by A and B, concept descriptions are denoted by C and D, and role names are denoted by r and s, all possibly with subscripts.

A *terminology* or TBox $\mathcal{T}$ is a finite set of (possibly primitive) concept definitions and role hierarchy axioms, whose syntax is an expression of the form $(A \sqsubseteq D)$ $A \equiv D$ and $r \sqsubseteq s$, respectively. The set $\mathsf{CN}^{\mathsf{def}}$ of *defined concept names* are concept names which appear on the left-hand side of a concept definition. Other concepts are called *primitive concept names*, denoted by $\mathsf{CN}^{\mathsf{pri}}$. A TBox $\mathcal{T}$ is called *unfoldable* if all concept definitions are *unique* and *acyclic* definitions. A concept definition A is unique if $\mathcal{T}$ contains at most one concept definition for $A \in \mathsf{CN}^{\mathsf{def}}$ and is acyclic if A is not referred directly or indirectly (via other concept definitions) to itself. For every primitive concept definition $A \sqsubseteq D$ in $\mathcal{T}$, each can be transformed into an equivalent

one by introducing a fresh concept name A' via the rule $A \sqsubseteq D \longrightarrow A \equiv A' \sqcap D$. When a TBox $\mathcal{T}$ is unfoldable, concept names can be expanded by exhaustively replacing all defined concept names by their definitions until only primitive concept names remain. Such concept names are called *fully expanded concept names.*[2]

Like primitive definitions, a role hierarchy axiom $r \sqsubseteq s$ can be transformed into a semantically equivalent role definition, by introducing a fresh role name r' via the similar rule $r \sqsubseteq s \longrightarrow r \equiv r' \sqcap s$. Role names occurring on the left-hand side of a role definition are called *defined role names*, collectively denoted by $\mathsf{RN}^{\mathsf{def}}$. All others are called *primitive role names*, collectively denoted by $\mathsf{RN}^{\mathsf{pri}}$. A set of all r's super roles, denoted by $\mathcal{R}_r$, is defined as $\mathcal{R}_r = \{s \in \mathsf{RN} | r \sqsubseteq^* s\}$ and, $r \sqsubseteq^* s$ if $r = s$ or $r_i \sqsubseteq r_{i+1} \in \mathcal{T}$ where $1 \leq i \leq n, r_1 = r, r_n = s$, and $*$ is a transitive closure.

An interpretation $\mathcal{I}$ is a pair $\mathcal{I} = \langle \Delta^{\mathcal{I}}, \cdot^{\mathcal{I}} \rangle$ where $\Delta^{\mathcal{I}}$ is a non-empty set representing the domain of the interpretation and $\cdot^{\mathcal{I}}$ is an interpretation function which assigns to every concept name A a set $A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$ and to every role name r a binary relation $r^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$. The interpretation function $\cdot^{\mathcal{I}}$ is inductively extended to $\mathcal{ELH}$ concepts in the usual manner:

$$\top^{\mathcal{I}} = \Delta^{\mathcal{I}}; \qquad (C \sqcap D)^{\mathcal{I}} = C^{\mathcal{I}} \cap D^{\mathcal{I}};$$

$$(\exists r.C)^{\mathcal{I}} = \left\{ a \in \Delta^{\mathcal{I}} \mid \exists b \in \Delta^{\mathcal{I}} : (a, b) \in r^{\mathcal{I}} \wedge b \in C^{\mathcal{I}} \right\}.$$

An interpretation $\mathcal{I}$ is said to be a *model* of a TBox $\mathcal{T}$ (in symbols, $\mathcal{I} \models \mathcal{T}$) if it satisfies all axioms in $\mathcal{T}$. $\mathcal{I}$ satisfies axioms $A \sqsubseteq C$, $A \equiv C$, and $r \sqsubseteq s$, respectively, if $A^{\mathcal{I}} \subseteq C^{\mathcal{I}}$, $A^{\mathcal{I}} = C^{\mathcal{I}}$, and $r^{\mathcal{I}} \subseteq s^{\mathcal{I}}$. The main inference problem for $\mathcal{ELH}$ is the subsumption problem. That is, given $C, D \in \mathsf{Con}(\mathcal{ELH})$ and a TBox $\mathcal{T}$, C is *subsumed* by D w.r.t. $\mathcal{T}$ (in symbols, $C \sqsubseteq_{\mathcal{T}} D$) if $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$ for every model $\mathcal{I}$ of $\mathcal{T}$. Furthermore, C and D are *equivalent* w.r.t. $\mathcal{T}$ (in symbols, $C \equiv_{\mathcal{T}} D$) if $C \sqsubseteq_{\mathcal{T}} D$ and $D \sqsubseteq_{\mathcal{T}} C$. When a TBox $\mathcal{T}$ is empty or is clear from the context, we omit to denote $\mathcal{T}$, i.e. $C \sqsubseteq D$ or $C \equiv D$.

Let $C \in \mathsf{Con}(\mathcal{ELH})$ be a fully expanded concept to the form: $P_1 \sqcap \cdots \sqcap P_m \sqcap \exists r_1.C_1 \sqcap \cdots \sqcap \exists r_n.C_n$, where $P_i \in \mathsf{CN}^{\mathsf{pri}}$, $r_j \in \mathsf{RN}$, $C_j \in \mathsf{Con}(\mathcal{ELH})$ in the same format, $1 \leq i \leq m$, and $1 \leq j \leq n$. The set $P_1, \ldots, P_m$ and the set $\exists r_1.C_1, \ldots, \exists r_n.C_n$ are denoted by $\mathcal{P}_C$ and $\mathcal{E}_C$, respectively, i.e. $\mathcal{P}_C = \{P_1, \ldots, P_m\}$ and $\mathcal{E}_C = \{\exists r_1.C_1, \ldots, \exists r_n.C_n\}$. An $\mathcal{ELH}$ concept description can be structurally transformed into the corresponding $\mathcal{ELH}$ description tree. The root v_0 of the $\mathcal{ELH}$ description tree $\mathcal{T}_C$ has $\{P_1, \ldots, P_m\}$ as its label and has n outgoing edges, each labeled with $\mathcal{R}_{r_j}$ to a vertex v_j for $1 \leq j \leq n$. Then, a subtree with the root v_j is defined recursively relative to the concept C_j. In [11, 12], a characterization of subsumption for the DL $\mathcal{ELH}$ w.r.t. an unfoldable TBox is proposed. Instead of considering concept descriptions, the so-called $\mathcal{ELH}$ description trees corresponding to those concept descriptions are considered. The subsumption is then characterized by an existence of a homomorphism in the reverse direction (cf. Theorem 1).

[2] In this work, we assume that concept names are fully expanded and the TBox can be omitted.

Definition 1 (Homomorphism [11, 12])**.** An $\mathcal{ELH}$ description tree $\mathcal{T}$ is a quintuple (V, E, rt, l, p) where V is a set of vertices, $E \subseteq V \times V$ is a set of edges, rt is the root, $l : V \to 2^{\mathsf{CN}^{\mathsf{pri}}}$ is a vertex labeling function, and $\rho : E \to 2^{\mathsf{RN}}$ is an edge labeling function. Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two $\mathcal{ELH}$ description trees, $v_1 \in V_1$ and $v_2 \in V_2$, there exists a homomorphism h from $\mathcal{T}_1$ to $\mathcal{T}_2$ (written as $h : \mathcal{T}_1 \to \mathcal{T}_2$) iff the following conditions are satisfied:

- $h(rt_1) = rt_2$ and $l_1(v_1) \subseteq l_2(h(v_1))$; and
- for each successor w_1 of v_1 in $\mathcal{T}_1$, $h(w_1)$ is a successor of $h(v_1)$ with $\rho_1(v_1, w_1) \subseteq \rho_2(h(v_1), h(w_1))$.

Example 2. (Continuation of Example 1) Each primitive definition can be transformed to a corresponding equivalent full definition as follows.

$$\begin{aligned} \mathsf{ActivePlace} &\equiv X \sqcap \mathsf{Place} \sqcap \exists \mathsf{canWalk}.\mathsf{Trekking} \sqcap \exists \mathsf{canSail}.\mathsf{Kayaking} \\ \mathsf{Mangrove} &\equiv Y \sqcap \mathsf{Place} \sqcap \exists \mathsf{canWalk}.\mathsf{Trekking} \\ \mathsf{Beach} &\equiv Z \sqcap \mathsf{Place} \sqcap \exists \mathsf{canSail}.\mathsf{Kayaking} \end{aligned}$$

where X, Y, and Z are fresh primitive concept names. Similarly, $\mathsf{canWalk} \equiv t \sqcap \mathsf{canMoveWithLegs}$ and $\mathsf{canSail} \equiv u \sqcap \mathsf{canTravelWithSails}$, where t and u are fresh primitive role names. In other words, $\mathcal{R}_{\mathsf{canWalk}} = \{t, \mathsf{canMoveWithLegs}\}$ and $\mathcal{R}_{\mathsf{canSail}} = \{u, \mathsf{canTravelWithSails}\}$. Figure 1 depicts $\mathcal{T}_{\mathsf{ActivePlace}}$, as an illustration.

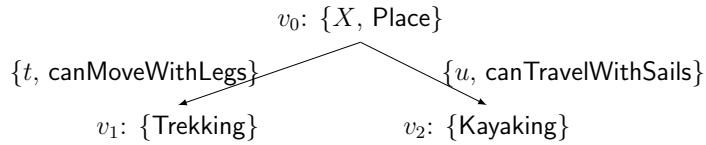


Figure 1. The description tree of concept ActivePlace

Theorem 1 ([11, 12])**.** Let $C, D \in \mathsf{Con}(\mathcal{ELH})$ and $\mathcal{T}_C$ and $\mathcal{T}_D$ be the corresponding description trees. Then, $C \sqsubseteq D$ iff there exists a homomorphism $h : \mathcal{T}_D \to \mathcal{T}_C$ that maps the root of $\mathcal{T}_D$ to the root of $\mathcal{T}_C$.

From Example 2, it is also not difficult to find a failed attempt of identifying a homomorphism mapping the root of $\mathcal{T}_{\mathsf{ActivePlace}}$ to the root of $\mathcal{T}_{\mathsf{Mangrove}}$, i.e. $h : \mathcal{T}_{\mathsf{ActivePlace}} \not\to \mathcal{T}_{\mathsf{Mangrove}}$. Hence, this infers $\mathsf{Mangrove} \not\sqsubseteq \mathsf{ActivePlace}$ by Theorem 1.

2.2 Concept Similarity Measure in DLs

Concept similarity measure (abbreviated as CSM) is a function mapping from a concept pair to a unit interval ($0 \leq x \leq 1$ where x is a real number). The higher the

value is mapped to, the more likely similarity of that concept pair may hold. In the following, we have formally defined the notion of CSM in DLs.

Definition 2. Given two concept descriptions $C, D \in \mathsf{Con}(\mathcal{L})$, a *concept similarity measure* w.r.t. a TBox $\mathcal{T}$ is a function $\sim_{\mathcal{T}} : \mathsf{Con}(\mathcal{L}) \times \mathsf{Con}(\mathcal{L}) \to [0,1]$ such that $C \sim_{\mathcal{T}} D = 1$ iff $C \equiv_{\mathcal{T}} D$ (*total similarity*) and $C \sim_{\mathcal{T}} D = 0$ indicates *total dissimilarity* between C and D.

When a TBox $\mathcal{T}$ is clear from the context, we simply write $\sim$. Furthermore, to avoid confusion on the symbols, $\sim_{\mathcal{T}}$ is used when referring to arbitrary measures.

The measure sim [13, 10] extends Theorem 1 to the case where no such homomorphism exists but there is some commonality. Since an extension to sim is presented in Subsection 4.1 for taking into account the agent's preferences, the original definitions of homomorphism degree hd and sim are included here for self-containment.

Definition 3 (Homomorphism Degree [10])**.** Let $\mathbf{T}^{\mathcal{ELH}}$ be a set of all $\mathcal{ELH}$ description trees and $\mathcal{T}_C, \mathcal{T}_D \in \mathbf{T}^{\mathcal{ELH}}$ correspond to two $\mathcal{ELH}$ concept names C and D, respectively. The *homomorphism degree* function hd: $\mathbf{T}^{\mathcal{ELH}} \times \mathbf{T}^{\mathcal{ELH}} \to [0,1]$ is inductively defined as follows:

$$\mathsf{hd}(\mathcal{T}_D, \mathcal{T}_C) = \mu \cdot \mathsf{p\text{-}hd}(\mathcal{P}_D, \mathcal{P}_C) + (1-\mu) \cdot \mathsf{e\text{-}set\text{-}hd}(\mathcal{E}_D, \mathcal{E}_C) \tag{1}$$

where $\mu = \frac{|\mathcal{P}_D|}{|\mathcal{P}_D \cup \mathcal{E}_D|}$ and $|\cdot|$ represents the set cardinality;

$$\mathsf{p\text{-}hd}(\mathcal{P}_D, \mathcal{P}_C) = \begin{cases} 1, & \text{if } \mathcal{P}_D = \emptyset, \\ \frac{|\mathcal{P}_D \cap \mathcal{P}_C|}{|\mathcal{P}_D|}, & \text{otherwise,} \end{cases} \tag{2}$$

$$\mathsf{e\text{-}set\text{-}hd}(\mathcal{E}_D, \mathcal{E}_C) = \begin{cases} 1, & \text{if } \mathcal{E}_D = \emptyset, \\ 0, & \text{if } \mathcal{E}_D \neq \emptyset \text{ and } \mathcal{E}_C = \emptyset, \\ \sum_{\epsilon_i \in \mathcal{E}_D} \frac{\max_{\epsilon_j \in \mathcal{E}_C} \{\mathsf{e\text{-}hd}(\epsilon_i, \epsilon_j)\}}{|\mathcal{E}_D|}, & \text{otherwise} \end{cases} \tag{3}$$

with ϵ_i, ϵ_j existential restrictions; and

$$\mathsf{e\text{-}hd}(\exists r.X, \exists s.Y) = \gamma(\nu + (1-\nu) \cdot \mathsf{hd}(\mathcal{T}_X, \mathcal{T}_Y)) \tag{4}$$

where $\gamma = \frac{|\mathcal{R}_r \cap \mathcal{R}_s|}{|\mathcal{R}_r|}$ and $0 \leq \nu < 1$.

The value of ν in Equation (4) determines how important the roles are to be considered for similarity between two existential restriction information. For instance, $\exists$canWalk.Trekking and $\exists$canWalk.Parading for dissimilar nested concepts Trekking and Parading should not be regarded as entirely dissimilar themselves. If ν is assigned the values 0.3, 0.4, 0.5, then $\mathsf{e\text{-}hd}(\exists\mathsf{canWalk.Trekking}, \exists\mathsf{canWalk.Parading})$ is 0.3, 0.4, 0.5, respectively. This value might vary among applications. In this work, ν is set to 0.4 for exemplifying the calculation of hd.

Theorem 2 ([10])**.** Let $C, D \in \mathsf{Con}(\mathcal{ELH})$ and $\mathcal{T}_C, \mathcal{T}_D$ be their corresponding description tree, respectively. Then, the following are equivalent:

1. $C \sqsubseteq D$; and
2. $\mathsf{hd}(\mathcal{T}_D, \mathcal{T}_C) = 1$.

Using a proof by induction, together with Theorem 1, it is not difficult to observe the correspondence between the homomorphism degree hd and subsumption. Intuitively, Theorem 2 describes a property of concept subsumption, i.e. C is a subconcept of D if the homomorphism degree of the corresponding description tree $\mathcal{T}_D$ to $\mathcal{T}_C$ is equal to 1, and vice versa.

Definition 4 ($\mathcal{ELH}$ Similarity Degree [10])**.** Let C and D be $\mathcal{ELH}$ concept names and $\mathcal{T}_C, \mathcal{T}_D$ be the corresponding description trees. Then, the $\mathcal{ELH}$ *similarity degree* between C and D (in symbols, $\mathsf{sim}(C, D)$) is defined as follows:

$$\mathsf{sim}(C, D) = \frac{\mathsf{hd}(\mathcal{T}_C, \mathcal{T}_D) + \mathsf{hd}(\mathcal{T}_D, \mathcal{T}_C)}{2}. \tag{5}$$

Example 3. (Continuation of Example 2)
For brevity, let ActivePlace, Mangrove, Beach, Place, Trekking, Kayaking, canWalk, and canSail be abbreviated as AP, M, B, P, T, K, cW, and cS, respectively. Using Definition 3, the homomorphism degree from $\mathcal{T}_{\mathsf{AP}}$ to $\mathcal{T}_{\mathsf{M}}$, or

$$\begin{aligned}\mathsf{hd}(\mathcal{T}_{\mathsf{AP}}, \mathcal{T}_{\mathsf{M}}) &= \left(\frac{2}{4}\right)\left(\frac{1}{2}\right) \\ &\quad + \left(\frac{2}{4}\right)\left(\frac{\max\{\mathsf{e\text{-}hd}(\exists \mathsf{cW.T}, \exists \mathsf{cW.T})\}}{2} + \frac{\max\{\mathsf{e\text{-}hd}(\exists \mathsf{cS.K}, \exists \mathsf{cW.T})\}}{2}\right) \\ &= \left(\frac{2}{4}\right)\left(\frac{1}{2}\right) + \left(\frac{2}{4}\right)\left(\frac{1+0}{2}\right) = 0.5.\end{aligned}$$

Similarly, $\mathsf{hd}(\mathcal{T}_{\mathsf{M}}, \mathcal{T}_{\mathsf{AP}}) = 0.67$, $\mathsf{hd}(\mathcal{T}_{\mathsf{AP}}, \mathcal{T}_{\mathsf{B}}) = 0.5$, and $\mathsf{hd}(\mathcal{T}_{\mathsf{B}}, \mathcal{T}_{\mathsf{AP}}) = 0.67$. Thus, $\mathsf{sim}(\mathsf{M}, \mathsf{AP}) = 0.59$ and $\mathsf{sim}(\mathsf{B}, \mathsf{AP}) = 0.59$

3 PREFERENCE PROFILE

We first introduced *preference profile* (denoted by π) in [14] as a collection of preferential elements in which the development of CSM should be considered. Its first intuition is to model different forms of preferences (of an agent) based on concept names and role names. Measures adopted this notion are flexible to be tuned by an agent and can determine the similarity conformable to that agent's perception.

The syntax and semantics of each form are given in term of partial functions because agents may not have preferences over all concept names and role names. We recommend to devise similarity measures with considerations on preference profile

if we aim at developing concept similarity measure for general purposes – a measure based on both subjective and objective factors. Mathematical definitions for each form of preferences are formally defined as follows.

Definition 5 (Primitive Concept Importance)**.** Let $\mathsf{CN}^{\mathsf{pri}}(\mathcal{T})$ be a set of primitive concept names occurring in $\mathcal{T}$. Then, a *primitive concept importance* is a partial function $\mathfrak{i}^{\mathfrak{c}} : \mathsf{CN} \to [0, 2]$[3], where $\mathsf{CN} \subseteq \mathsf{CN}^{\mathsf{pri}}(\mathcal{T})$.

For any $A \in \mathsf{CN}^{\mathsf{pri}}(\mathcal{T})$, $\mathfrak{i}^{\mathfrak{c}}(A) = 1$ captures an expression of normal importance for A, $\mathfrak{i}^{\mathfrak{c}}(A) > 1$ (and $\mathfrak{i}^{\mathfrak{c}}(A) < 1$) indicates that A has higher (and lower, respectively) importance, and $\mathfrak{i}^{\mathfrak{c}}(A) = 0$ indicates that A is of no importance to the agent.

Example 4. (Continuation of Example 2) Suppose that an agent A is using a similarity measure for querying some names similar to $\mathsf{ActivePlace}$. He concerns that those names will be similar to $\mathsf{ActivePlace}$ if they are *places*. Thus, the agent can express this preference as $\mathfrak{i}^{\mathfrak{c}}(\mathsf{Place}) = 2$, i.e., values should be higher than 1.

On the other hand, suppose he does not care if those are *places* or not, he may express this preference as $\mathfrak{i}^{\mathfrak{c}}(\mathsf{Place}) = 0$, i.e., values must be equal to 0.

Definition 6 (Role Importance)**.** Let $\mathsf{RN}(\mathcal{T})$ be a set of role names occurring in $\mathcal{T}$. Then, a *role importance* is a partial function $\mathfrak{i}^{\mathfrak{r}} : \mathsf{RN} \to [0, 2]$, where $\mathsf{RN} \subseteq \mathsf{RN}(\mathcal{T})$.

For any $r \in \mathsf{RN}(\mathcal{T})$, $\mathfrak{i}^{\mathfrak{r}}(r) = 1$ captures an expression of normal importance for r, $\mathfrak{i}^{\mathfrak{r}}(r) > 1$ (and $\mathfrak{i}^{\mathfrak{r}}(r) < 1$) indicates that r has higher (and lower, respectively) importance, and $\mathfrak{i}^{\mathfrak{r}}(r) = 0$ indicates that r is of no importance to the agent.

Example 5 (Continuation of Example 2)**.** Suppose that the agent A wants to enjoy *walking*. He may express this preference as $\mathfrak{i}^{\mathfrak{r}}(\mathsf{canWalk}) = 2$, i.e., values should be higher than 1.

Definition 7 (Primitive Concepts Similarity)**.** Let $\mathsf{CN}^{\mathsf{pri}}(\mathcal{T})$ be a set of primitive concept names occurring in $\mathcal{T}$. For $A, B \in \mathsf{CN}^{\mathsf{pri}}(\mathcal{T})$, a *primitive concepts similarity* is a partial function $\mathfrak{s}^{\mathfrak{c}} : \mathsf{CN} \times \mathsf{CN} \to [0, 1]$, where $\mathsf{CN} \subseteq \mathsf{CN}^{\mathsf{pri}}(\mathcal{T})$, such that $\mathfrak{s}^{\mathfrak{c}}(A, B) = \mathfrak{s}^{\mathfrak{c}}(B, A)$ and $\mathfrak{s}^{\mathfrak{c}}(A, A) = 1$.

For $A, B \in \mathsf{CN}^{\mathsf{pri}}(\mathcal{T})$, $\mathfrak{s}^{\mathfrak{c}}(A, B) = 1$ captures an expression of total similarity between A and B and $\mathfrak{s}^{\mathfrak{c}}(A, B) = 0$ captures an expression of their total dissimilarity.

Example 6 (Continuation of Example 2)**.** Suppose that the agent A believes that *trekking* and *kayaking* invoke similar feeling. Thus, he can express $\mathfrak{s}^{\mathfrak{c}}(\mathsf{Trekking}, \mathsf{Kayaking}) = 0.1$, i.e., values should be higher than 0.

[3] In the original definition of preference profile, elements in the domains of both $\mathfrak{i}^{\mathfrak{c}}$ and $\mathfrak{i}^{\mathfrak{r}}$ are mapped to $\mathbb{R}_{\geq 0}$, which is a minor error.

Another example is the similarity of concepts Pet_1 and Pet_2, in which both are defined as follows: $\mathsf{Pet}_1 \sqsubseteq \mathsf{Dog} \sqcap \exists \mathsf{hasOwned}.\mathsf{Human}$; $\mathsf{Pet}_2 \sqsubseteq \mathsf{Cat} \sqcap \exists \mathsf{hasOwned}.\mathsf{Human}$. Here, Dog and Cat are both primitive concept names. Intuitively, Dog and Cat are similar, then we may attach this knowledge in form of $\mathfrak{s}^{\mathsf{c}}$ in order to yield more accuracy on the measure.

Definition 8 (Primitive Roles Similarity)**.** Let $\mathsf{RN}^{\mathsf{pri}}(\mathcal{T})$ be a set of primitive role names occurring in $\mathcal{T}$. For $r, s \in \mathsf{RN}^{\mathsf{pri}}(\mathcal{T})$, a *primitive roles similarity* is a partial function $\mathfrak{s}^{\mathsf{r}} : \mathsf{RN} \times \mathsf{RN} \to [0, 1]$, where $\mathsf{RN} \subseteq \mathsf{RN}^{\mathsf{pri}}(\mathcal{T})$, such that $\mathfrak{s}^{\mathsf{r}}(r, s) = \mathfrak{s}^{\mathsf{r}}(s, r)$ and $\mathfrak{s}^{\mathsf{r}}(r, r) = 1$.

For $r, s \in \mathsf{RN}(\mathcal{T})$, $\mathfrak{s}^{\mathsf{r}}(r, s) = 1$ captures an expression of total similarity between r and s and $\mathfrak{s}^{\mathsf{r}}(r, s) = 0$ captures an expression of their total dissimilarity.

Example 7 (Continuation of Example 2)**.** Suppose that the agent A believes that *moving with legs* and *traveling with sails* invoke similar feeling. He may express $\mathfrak{s}^{\mathsf{r}}(\mathsf{canMoveWithLegs}, \mathsf{canTravelWithSails}) = 0.1$, i.e., values should be higher than 0.

Basically, our motivations of both functions $\mathfrak{s}^{\mathsf{c}}$ and $\mathfrak{s}^{\mathsf{r}}$ are the same, i.e., we aim at attaching subjective feeling of proximity (about primitive concept names and primitive role names) into a measure. In DLs, different primitive concept names (and also primitive role names) are considered to be total dissimilarity even though they may be recognized as being similar in real-world domains.

Definition 9 (Role Discount Factor)**.** Let $\mathsf{RN}(\mathcal{T})$ be a set of role names occurring in $\mathcal{T}$. Then, a *role discount factor* is a partial function $\mathfrak{d} : \mathsf{RN} \to [0, 1]$, where $\mathsf{RN} \subseteq \mathsf{RN}(\mathcal{T})$.

For any $r \in \mathsf{RN}(\mathcal{T})$, $\mathfrak{d}(r) = 1$ captures an expression of total importance on the role (beyond a corresponding nested concept) and $\mathfrak{d}(r) = 0$ captures an expression of total importance on a nested concept (beyond the correspondent role r).

Example 8 (Continuation of Example 2)**.** Suppose that the agent A does not concern much if places permit to either walk or to sail. He would rather consider on actual activities which he can perform. Thus, he may express $\mathfrak{d}(\mathsf{canWalk}) = 0.3$ and $\mathfrak{d}(\mathsf{canSail}) = 0.3$, i.e., values should be close to 0.

Definition 10 (Preference Profile)**.** A *preference profile*, in symbol π, is a quintuple $\langle \mathfrak{i}^{\mathsf{c}}, \mathfrak{i}^{\mathsf{r}}, \mathfrak{s}^{\mathsf{c}}, \mathfrak{s}^{\mathsf{r}}, \mathfrak{d} \rangle$ where $\mathfrak{i}^{\mathsf{c}}, \mathfrak{i}^{\mathsf{r}}, \mathfrak{s}^{\mathsf{c}}, \mathfrak{s}^{\mathsf{r}}$, and $\mathfrak{d}$ are as defined above and the *default preference profile*, in symbol π_0, is the quintuple $\langle \mathfrak{i}^{\mathsf{c}}_0, \mathfrak{i}^{\mathsf{r}}_0, \mathfrak{s}^{\mathsf{c}}_0, \mathfrak{s}^{\mathsf{r}}_0, \mathfrak{d}_0 \rangle$ where

$$\mathfrak{i}^{\mathfrak{c}}_0(A) = 1 \text{ for all } A \in \mathsf{CN}^{\mathsf{pri}}(\mathcal{T}),$$
$$\mathfrak{i}^{\mathfrak{r}}_0(r) = 1 \text{ for all } r \in \mathsf{RN}(\mathcal{T}),$$
$$\mathfrak{s}^{\mathfrak{c}}_0(A, B) = 0 \text{ for all } (A, B) \in \mathsf{CN}^{\mathsf{pri}}(\mathcal{T}) \times \mathsf{CN}^{\mathsf{pri}}(\mathcal{T}),$$
$$\mathfrak{s}^{\mathfrak{r}}_0(r, s) = 0 \text{ for all } (r, s) \in \mathsf{RN}^{\mathsf{pri}}(\mathcal{T}) \times \mathsf{RN}^{\mathsf{pri}}(\mathcal{T}), \text{ and}$$
$$\mathfrak{d}_0(r) = 0.4 \text{ for all } r \in \mathsf{RN}(\mathcal{T}).$$

Intuitively, the default preference profile π_0 represents the agent's preference in the default manner, i.e., when preferences are not given. That is, every $A \in \mathsf{CN}^{\mathsf{pri}}$ has normal importance and so does every $r \in \mathsf{RN}$. Also, every $(A, B) \in \mathsf{CN}^{\mathsf{pri}} \times \mathsf{CN}^{\mathsf{pri}}$ is totally different and so does every $(r, s) \in \mathsf{RN}^{\mathsf{pri}} \times \mathsf{RN}^{\mathsf{pri}}$. Lastly, every $r \in \mathsf{RN}$ is considered 0.4 importance for the similarity of two existential restriction information. It is interesting to note that changes in the definition of the default preference profile yield different interpretations of the default preference and thereby may produce a different degree of similarity under the default manner. As for its exemplification, the value 0.4 is used by $\mathfrak{d}_0$ to conform with the value of ν used in this work.

In this work, a preference profile of an agent is denoted by subscribing that agent below π, e.g., π_{A} represents a preference profile of the agent A.

4 SIMILARITY MEASURE UNDER PREFERENCE PROFILE

A numerical value determined by CSM indicates a degree of similarity of two concept descriptions w.r.t. the sole objective aspects. That is, either their structures or their interpretations are similar (cf. Section 7). For example, $\mathsf{sim}(\mathsf{ActivePlace}, \mathsf{Mangrove}) = 0.59$ and $\mathsf{sim}(\mathsf{ActivePlace}, \mathsf{Beach}) = 0.59$ indicates that the similarity of ActivePlace and Mangrove, and that of ActivePlace and Beach are equivalently 59 %. However, this information may not be useful for the agent to make decisions.

In this section, we present a conceptual notion for *concept similarity measure under the agent's preferences* (originally introduced in [15]) and its desirable properties. We also present the measure sim^{π} by adopting preference profile onto the measure sim. Our first intuition is to exemplify the applicability of preference profile onto an arbitrary existing measure. This shows that our proposed notion of preference profile can be considered as a collection of noteworthy aspects for the development of concept similarity measure under the agent's preferences.

Definition 11. Given a preference profile π, two concepts $C, D \in \mathsf{Con}(\mathcal{L})$, and a TBox $\mathcal{T}$, a *concept similarity measure under preference profile* w.r.t. a TBox $\mathcal{T}$ is a function $\overset{\pi}{\sim}_{\mathcal{T}}\colon \mathsf{Con}(\mathcal{L}) \times \mathsf{Con}(\mathcal{L}) \to [0, 1]$.

When a TBox $\mathcal{T}$ is clear from the context, we simply write $\overset{\pi}{\sim}$. Furthermore, to avoid confusion on the symbols, $\overset{\pi}{\sim}_{\mathcal{T}}$ is used when referring to arbitrary measures.

The notion $\overset{\pi}{\sim}$ may be informally read as *the computation of* $\sim$ *is influenced by* π. That informal interpretation shapes our intuition to consider this kind as a generalization of CSM in DLs. With adopting of this viewpoint of the interpretation, we can agree that sim^{π} (Subsection 4.1) is informally interpreted as *we compute* sim *(Definition 4) under an existence of a given* π.

Basically, the notion $\overset{\pi}{\sim}$ is a function mapping a pair of two concept descriptions w.r.t. a particular π to a unit interval. We identify a property called *preference invariance w.r.t. equivalence* in our preliminary study [15]. Now, we aim at identifying more important properties of $\overset{\pi}{\sim}$. We start by investigating important properties of CSM existing in the literature (e.g. [16, 9]). Our primary motivation is to identify CSM's properties which are also reasonable for $\overset{\pi}{\sim}$. The following collects fundamental properties for the introduced concept similarity measure under preference profile. They can be used to answer the question *What could be good preference-based similarity measures?* In other words, any preference-based measures satisfying the fundamental properties are considered to be good ones.

Formally, let $C, D, E \in \mathsf{Con}(\mathcal{L})$ and Π be a countably infinite set of preference profile. Then, we call a concept similarity measure under preference profile $\overset{\pi}{\sim}$ is:

1. *symmetric* iff $\forall \pi' \in \Pi : (C \overset{\pi'}{\sim} D = D \overset{\pi'}{\sim} C)$;
2. *equivalence invariant* iff $C \equiv D \Longrightarrow \forall \pi' \in \Pi : (C \overset{\pi'}{\sim} E = D \overset{\pi'}{\sim} E)$;
3. *structurally dependent* iff for any finite sets of concepts C_1 and C_2 with the following conditions:

   - $\mathsf{C}_1 \subseteq \mathsf{C}_2$,
   - concepts $A, B \notin \mathsf{C}_2$,
   - $\mathfrak{i}^{\mathfrak{c}}(\Phi) > 0$ if Φ is primitive and $\Phi \in \mathsf{C}_2$, and
   - $\mathfrak{i}^{\mathfrak{r}}(\varphi) > 0$ if Φ is existential, i.e. $\Phi \coloneqq \exists \varphi.\Psi$, and $\Phi \in \mathsf{C}_2$,

   the concepts $C \coloneqq \bigsqcap(\mathsf{C}_1 \cup \{A\})$, $D \coloneqq \bigsqcap(\mathsf{C}_1 \cup \{B\})$, $E \coloneqq \bigsqcap(\mathsf{C}_2 \cup \{A\})$ and $F \coloneqq \bigsqcap(\mathsf{C}_2 \cup \{B\})$ fulfill the condition $\forall \pi' \in \Pi : (C \overset{\pi'}{\sim} D \leq E \overset{\pi'}{\sim} F)$; and
4. *preference invariant w.r.t. equivalence* iff $C \equiv D \Longleftrightarrow \forall \pi' \in \Pi : C \overset{\pi'}{\sim} D = 1$.

Next, we discuss the underlying intuitions of each property subsequently. We note that the properties 1 to 3 are adopted from [16, 9]. However, to the best of our knowledge, the property 4 is first time introduced for *concept similarity measure under preference profile* in this work (originally introduced in [15]).

Let Π be a countably infinite set of preference profile. In the following, we discuss the intuitive interpretation of each property. Firstly, *symmetry* states that an order of concepts in question does not influence the notion $\overset{\pi}{\sim}$ for any $\pi \in \Pi$. For instance, $\mathsf{Mangrove} \overset{\pi}{\sim} \mathsf{Beach} = \mathsf{Beach} \overset{\pi}{\sim} \mathsf{Mangrove}$. This property is controversial as cognitive sciences believes that similarity is asymmetric. However, substantial work in DLs [16, 10, 17, 6, 8, 9, 7, 15, 5] prefers symmetry (merely [4, 18] prefer asymmetry). Here, we also agree on the symmetry because axiomatic information

in TBox is not dynamically changed. Furthermore, the notion of preference profile studied in this work is also static, i.e., it can be changed merely by tuning.

Secondly, *equivalence invariance* (alternatively called *equivalence soundness* [9] in the context of dissimilarity measure) states that if two concepts C and D are logically equivalent, then measuring the similarity of each toward the third concept E w.r.t. any $\pi \in \Pi$ must be the same. For instance, let $C \equiv \exists \mathsf{canWalk}.\mathsf{Trekking}$ and $D \equiv \exists \mathsf{canWalk}.\mathsf{Trekking}$. It is clear that C and D are logically equivalent. Therefore, let $E \in \mathsf{Con}(\mathcal{L})$, $C \overset{\pi}{\sim} E = D \overset{\pi}{\sim} E$ for any $\pi \in \Pi$.

Thirdly, the notion of *structural dependence* is originally introduced by Tversky in [19]. Later, the authors of [16] collect it as another important properties for CSM in their work. Basically, in Tversky's model, an object is considered as a set of features. Then, the similarity of two objects is measured by the relationship between a number of common features and a number of different features. Extending this idea to $\overset{\pi}{\sim}$ gives the meaning that the similarity of two concepts C, D increases if a more number of *equivalent* concepts is shared and each is considered *important*.

Lastly, *preference invariance w.r.t. equivalence* states that if two concepts are logically equivalent, then the similarity degree of two concepts under preference profile π is always one for every $\pi \in \Pi$, and vice versa. Taking the negation both sides, this means $C \not\equiv D \iff \exists \pi \in \Pi : C \overset{\pi}{\sim} D \neq 1$. For instance, let $C \equiv \exists \mathsf{canWalk}.\mathsf{Trekking}$ and $D \equiv \exists \mathsf{canWalk}.\mathsf{Parading}$. It is clear that C and D are not logically equivalent, then taking $\pi = \pi_0$ obtains $C \overset{\pi_0}{\sim} D \neq 1$; though, taking $\pi = \pi_1$ where $\mathfrak{s}^{\mathfrak{c}}(\mathsf{Trekking}, \mathsf{Parading}) = 1$ is defined in π_1 yields $C \overset{\pi_1}{\sim} D = 1$.

There are several properties which are not considered as fundamental properties of concept similarity measure under preference profile because the behaviors may not obey their properties when used under *non-default* preference profiles, e.g. *reverse subsumption preserving*. In [16], a measure $\sim$ satisfies the *reverse subsumption preserving* iff, for any concepts C, D, and E, $C \sqsubseteq D \sqsubseteq E \Longrightarrow C \sim E \leq D \sim E$. The property states that the similarity of D and E is higher than the one of C and E because E is closer to D than C. To refute it, we need only one preference profile π such that the implication does not hold (cf. Example 9), i.e., to show that $(C \sqsubseteq D \sqsubseteq E)$ and $\exists \pi \in \Pi : (C \overset{\pi}{\sim} E > D \overset{\pi}{\sim} E)$.

Example 9. Suppose concepts $\mathsf{A}_1, \mathsf{A}_2, \mathsf{A}_3$, and A_4 are primitive. Query describes features of an item that an agent is searching for. Item_1 and Item_2 are items, which compose of features $\mathsf{A}_1, \mathsf{A}_2, \mathsf{A}_3$ and $\mathsf{A}_1, \mathsf{A}_2, \mathsf{A}_3, \mathsf{A}_4$, respectively.

$$\mathsf{Query} \sqsubseteq \mathsf{A}_1 \sqcap \mathsf{A}_2$$

$$\mathsf{Item}_1 \sqsubseteq \mathsf{A}_1 \sqcap \mathsf{A}_2 \sqcap \mathsf{A}_3$$

$$\mathsf{Item}_2 \sqsubseteq \mathsf{A}_1 \sqcap \mathsf{A}_2 \sqcap \mathsf{A}_3 \sqcap \mathsf{A}_4$$

The ontology shows the hierarchy: $\mathsf{Item}_2 \sqsubseteq \mathsf{Item}_1 \sqsubseteq \mathsf{Query}$. By taking $\mathfrak{s}^{\mathfrak{c}}(\mathsf{A}_2, \mathsf{A}_4) = 1$, it is reasonable to conclude that $\mathsf{Item}_2 \overset{\pi}{\sim} \mathsf{Query} > \mathsf{Item}_1 \overset{\pi}{\sim} \mathsf{Query}$ due to an increased number of totally similar concepts.

Our proceeding paper [5] studies CSM for the DL $\mathcal{FL}_0$. In this paper, we suggest two measures, viz. the skeptical measure $\sim^s$ and the credulous measure $\sim^c$, which are derived from the known structural characterization subsumption through inclusion of regular languages. This fact exhibits that there is not a unique CSM for similarity-based applications. Which CSMs should be used depends on concrete applications, especially the type of a rational agent. For example, when employing the notion $\sim$ to a query answering system, a credulous agent may want to see answers as much as possible; hence, the measure $\sim^c$ is employed. On the other hand, a skeptical agent would like to see sufficient relevant answers; hence, the measure $\sim^s$ is employed. This idea is generalized and is extended toward the notion $\overset{\pi}{\sim}$ to be used under different agent's profiles. In essence, if an arbitrary concept similarity measure under preference profile $\overset{\pi}{\sim}$ is fixed, measuring the similarity of two concepts under different preference profiles may yield different values.

Definition 12. Let Π be a countably infinite set of preference profile and $\pi_1, \pi_2 \in \Pi$. For any fixed measure $\overset{\pi}{\sim}$, the concept similarity measure under π_1 is *more skeptical than* π_2 (denoted by $\overset{\pi_1}{\sim} \preceq \overset{\pi_2}{\sim}$) if $C \overset{\pi_1}{\sim} D \leq C \overset{\pi_2}{\sim} D$ for all $C, D \in \mathsf{Con}(\mathcal{L})$.

4.1 The Measure sim^π

To develop an instance of $\overset{\pi}{\sim}$, we generalize sim for all aspects of preference profile. As a result, the new measure sim^π is also driven by the structural subsumption characterization by means of tree homomorphism for the DL $\mathcal{ELH}$.

We start by presenting each aspect of preference profile in term of *total functions* in order to avoid computing on null values. A *total importance function* is firstly introduced as $\hat{\mathfrak{i}} : \mathsf{CN}^{\mathsf{pri}} \cup \mathsf{RN} \to [0, 2]$ based on the primitive concept importance and the role importance.

$$\hat{\mathfrak{i}}(x) = \begin{cases} \mathfrak{i}^{\mathfrak{c}}(x), & \text{if } x \in \mathsf{CN}^{\mathsf{pri}} \text{and } \mathfrak{i}^{\mathfrak{c}} \text{ is defined on } x, \\ \mathfrak{i}^{\mathfrak{r}}(x), & \text{if } x \in \mathsf{RN} \text{ and } \mathfrak{i}^{\mathfrak{r}} \text{ is defined on } x, \\ 1, & \text{otherwise.} \end{cases} \tag{6}$$

A *total similarity function* is also presented as $\hat{\mathfrak{s}} : (\mathsf{CN}^{\mathsf{pri}} \times \mathsf{CN}^{\mathsf{pri}}) \cup (\mathsf{RN}^{\mathsf{pri}} \times \mathsf{RN}^{\mathsf{pri}}) \to [0, 1]$ using the primitive concepts similarity and the primitive roles similarity.

$$\hat{\mathfrak{s}}(x, y) = \begin{cases} 1, & \text{if } x = y, \\ \mathfrak{s}^{\mathfrak{c}}(x, y), & \text{if } (x, y) \in \mathsf{CN}^{\mathsf{pri}} \times \mathsf{CN}^{\mathsf{pri}} \text{ and } \mathfrak{s}^{\mathfrak{c}} \text{ is defined on } (x, y), \\ \mathfrak{s}^{\mathfrak{r}}(x, y), & \text{if } (x, y) \in \mathsf{RN}^{\mathsf{pri}} \times \mathsf{RN}^{\mathsf{pri}} \text{ and } \mathfrak{s}^{\mathfrak{r}} \text{ is defined on } (x, y), \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Similarly, a *total role discount factor function*[4] is presented in the following in term of a function $\hat{\mathfrak{d}} : \mathsf{RN} \to [0,1]$ based on the role discount factor.

$$\hat{\mathfrak{d}}(x) = \begin{cases} \mathfrak{d}(x), & \text{if } \mathfrak{d} \text{ is defined on } x, \\ 0.4, & \text{otherwise.} \end{cases} \tag{8}$$

The next step is to generalize the notion of homomorphism degree hd (Definition 3). Let $C, D \in \mathsf{Con}(\mathcal{ELH})$ and $r, s \in \mathsf{RN}$. Also, let $\mathcal{T}_C, \mathcal{T}_D, \mathcal{P}_C, \mathcal{P}_D, \mathcal{E}_C, \mathcal{E}_D, \mathcal{R}_r$, and $\mathcal{R}_s$ be as defined in Subsection 2.2. The homomorphism degree under preference profile π from $\mathcal{T}_D$ to $\mathcal{T}_C$ can be formally defined in Definition 13.

Definition 13. Let $\mathbf{T}^{\mathcal{ELH}}$ be a set of all $\mathcal{ELH}$ description trees, and $\pi = \langle \mathfrak{i}^{\mathfrak{c}}, \mathfrak{i}^{\mathfrak{r}}, \mathfrak{s}^{\mathfrak{c}}, \mathfrak{s}^{\mathfrak{r}}, \mathfrak{d} \rangle$ be a preference profile. The *homomorphism degree under preference profile* π is a function $\mathsf{hd}^{\pi} : \mathbf{T}^{\mathcal{ELH}} \times \mathbf{T}^{\mathcal{ELH}} \to [0,1]$ defined inductively as follows:

$$\mathsf{hd}^{\pi}(\mathcal{T}_D, \mathcal{T}_C) = \mu^{\pi}(\mathcal{P}_D, \mathcal{E}_D) \cdot \mathsf{p\text{-}hd}^{\pi}(\mathcal{P}_D, \mathcal{P}_C) + (1 - \mu^{\pi}(\mathcal{P}_D, \mathcal{E}_D)) \cdot \mathsf{e\text{-}set\text{-}hd}^{\pi}(\mathcal{E}_D, \mathcal{E}_C) \tag{9}$$

where

$$\mu^{\pi}(\mathcal{P}_D, \mathcal{E}_D) = \begin{cases} 1, & \text{if } \sum_{A \in \mathcal{P}_D} \hat{\mathfrak{i}}(A) + \sum_{\exists r.X \in \mathcal{E}_D} \hat{\mathfrak{i}}(r) = 0, \\ \frac{\sum_{A \in \mathcal{P}_D} \hat{\mathfrak{i}}(A)}{\sum_{A \in \mathcal{P}_D} \hat{\mathfrak{i}}(A) + \sum_{\exists r.X \in \mathcal{E}_D} \hat{\mathfrak{i}}(r)}, & \text{otherwise;} \end{cases} \tag{10}$$

$$\mathsf{p\text{-}hd}^{\pi}(\mathcal{P}_D, \mathcal{P}_C) = \begin{cases} 1, & \text{if } \sum_{A \in \mathcal{P}_D} \hat{\mathfrak{i}}(A) = 0, \\ 0, & \text{if } \sum_{A \in \mathcal{P}_D} \hat{\mathfrak{i}}(A) \neq 0, \\ & \quad \text{and } \sum_{B \in \mathcal{P}_C} \hat{\mathfrak{i}}(B) = 0, \\ \frac{\sum_{A \in \mathcal{P}_D} \hat{\mathfrak{i}}(A) \cdot \max_{B \in \mathcal{P}_C} \{\hat{\mathfrak{s}}(A,B)\}}{\sum_{A \in \mathcal{P}_D} \hat{\mathfrak{i}}(A)}, & \text{otherwise;} \end{cases} \tag{11}$$

$$\mathsf{e\text{-}set\text{-}hd}^{\pi}(\mathcal{E}_D, \mathcal{E}_C) = \begin{cases} 1, & \text{if } \sum_{\exists r.X \in \mathcal{E}_D} \hat{\mathfrak{i}}(r) = 0, \\ 0, & \text{if } \sum_{\exists r.X \in \mathcal{E}_D} \hat{\mathfrak{i}}(r) \neq 0 \\ & \quad \text{and } \sum_{\exists s.Y \in \mathcal{E}_C} \hat{\mathfrak{i}}(s) = 0, \\ \frac{\sum_{\exists r.X \in \mathcal{E}_D} \hat{\mathfrak{i}}(r) \cdot \max_{\epsilon_j \in \mathcal{E}_C} \{\mathsf{e\text{-}hd}^{\pi}(\exists r.X, \epsilon_j)\}}{\sum_{\exists r.X \in \mathcal{E}_D} \hat{\mathfrak{i}}(r)}, & \text{otherwise;} \end{cases} \tag{12}$$

where ϵ_j is an existential restriction; and

$$\mathsf{e\text{-}hd}^{\pi}(\exists r.X, \exists s.Y) = \gamma^{\pi}(r,s) \cdot (\hat{\mathfrak{d}}(r) + (1 - \hat{\mathfrak{d}}(r)) \cdot \mathsf{hd}^{\pi}(\mathcal{T}_X, \mathcal{T}_Y)) \tag{13}$$

[4] We set the default value to 0.4 to comply with the default value of π_0.

where

$$\gamma^{\pi}(r,s) = \begin{cases} 1, & \text{if } \sum_{r' \in \mathcal{R}_r} \hat{\mathsf{i}}(r') = 0, \\ \frac{\sum_{r' \in \mathcal{R}_r} \hat{\mathsf{i}}(r') \cdot \max_{s' \in \mathcal{R}_s} \{\hat{\mathfrak{s}}(r',s')\}}{\sum_{r' \in \mathcal{R}_r} \hat{\mathsf{i}}(r')}, & \text{otherwise.} \end{cases} \tag{14}$$

Intuitively, the function hd^{π} (Equation (9)) is defined as the weighted sum of the degree under preferences of the vertex set commonalities ($\mathsf{p\text{-}hd}^{\pi}$) and the degree under preferences of edge condition matching ($\mathsf{e\text{-}set\text{-}hd}^{\pi}$). Equation (11) calculates the average of the best matching under preferences of primitive concepts in $\mathcal{P}_D$. Equation (13) calculates the degree under preferences of a potential homomorphism of a matching edge. If edge labels share some commonalities under preferences (Equation (14)), i.e. $0 < \gamma^{\pi} \leq 1$, then part of the edge matching is satisfied; but the successors' labels and structures have yet to be checked. This is defined recursively as $\mathsf{hd}^{\pi}(\mathcal{T}_X, \mathcal{T}_Y)$ in Equation (13). Equation (12) calculates the best possible edge matching under preferences of each edge in $\mathcal{E}_D$ and returns the average thereof.

The weight μ^{π} in Equation (9) determines how important the primitive concept names are to be considered for preference-based similarity. For the special case where $D = \top$, i.e. $\mathcal{P}_D = \mathcal{E}_D = \emptyset$, μ^{π} is irrelevant as $\mathcal{T}_{\top}$ is the smallest $\mathcal{ELH}$ description tree and $\mathsf{hd}^{\pi}(\mathcal{T}_{\top}, \mathcal{T}_C) = 1$ for all concepts C.

It is to be mentioned that the function hd^{π} may look similar to $simi_d$ as both are recursive definitions for the same DL $\mathcal{ELH}$. However, they are obviously different caused by the distinction of their inspirations and their viewpoints of the development. While hd^{π} is inspired by the homomorphism-based structural subsumption characterization, $simi_d$ is inspired by the Jaccard Index [20]. Technically speaking, $simi_d$ employs t-conorm instead of fixing an operator. However, unlike $simi_d$, the use of μ^{π} for determining how primitive concepts are weighted and the use of γ^{π} for determining the proportion of shared super roles are employed. Furthermore, $simi_d$ is originated from the viewpoint of CSM, thus some aspects of preference profile are missed; though some may exist. We continue the discussion in Section 7.

The function hd^{π} yields a numerical value that represents structural similarity w.r.t. a particular profile π of a concept against another concept. As both directions constitute the degree of two concepts being equivalent, the measure sim^{π} is also defined by means of these two directional computations.

Definition 14. Let $C, D \in \mathsf{Con}(\mathcal{ELH})$, $\mathcal{T}_C$ and $\mathcal{T}_D$ be the corresponding description trees, and $\pi = \langle \mathfrak{i}^{\mathfrak{c}}, \mathfrak{i}^{\mathfrak{r}}, \mathfrak{s}^{\mathfrak{c}}, \mathfrak{s}^{\mathfrak{r}}, \mathfrak{d} \rangle$ be a preference profile. Then, the $\mathcal{ELH}$ *similarity measure under preference profile* π between C and D (denoted by $\mathsf{sim}^{\pi}(C,D)$) is defined as follows:

$$\mathsf{sim}^{\pi}(C,D) = \frac{\mathsf{hd}^{\pi}(\mathcal{T}_C, \mathcal{T}_D) + \mathsf{hd}^{\pi}(\mathcal{T}_D, \mathcal{T}_C)}{2}. \tag{15}$$

Intuitively, the degree of similarity under a certain π is the average of the degree of having homomorphism under the same π in both directions. Note that ones may also argue to calculate the value by using alternative binary operators

accepting unit intervals, e.g. based on the multiplication (in symbols, $\mathsf{mul\text{-}sim}^{\pi}$) on both directions of hd^{π} or the root mean square (in symbols, $\mathsf{rms\text{-}sim}^{\pi}$) on values of both directions [13]. Unfortunately, those give unsatisfactory values for the extreme cases. For example, $\mathsf{mul\text{-}sim}^{\pi}(A, \top) = 0 \times 1 = 0$ and $\mathsf{rms\text{-}sim}^{\pi}(A, \top) = \sqrt{\frac{0^2+1^2}{2}} = 0.707$, whereas $\mathsf{sim}^{\pi}(A, \top) = \frac{0+1}{2} = 0.5$. Since $\mathsf{mul\text{-}sim}^{\pi}(C, D) \leq \mathsf{sim}^{\pi}(C, D) \leq \mathsf{rms\text{-}sim}^{\pi}(C, D)$ for any concepts C and D, we believe that the average-based definition given above is the most appropriate method. Based on this form, sim^{π} is basically considered as a generalization of sim which determines similarity under preference profile, i.e., behavioral expectation of the measure will conform to the agent's perception.

We present an example about the calculation of sim^{π} in the following.

Example 10. (Continuation of Example 1) Let enrich the example. Assume the agent A's preference profile is defined as follows: (i) $\mathfrak{i}^{\mathfrak{r}}(\mathsf{Place}) = 2$; (ii) $\mathfrak{i}^{\mathfrak{r}}(\mathsf{canWalk}) = 2$; (iii) $\mathfrak{s}^{\mathfrak{r}}(\mathsf{Trekking}, \mathsf{Kayaking}) = 0.1$; (iv) $\mathfrak{s}^{\mathfrak{r}}(\mathsf{canMoveWithLegs}, \mathsf{canTravelWithSails}) = 0.1$; (v) $\mathfrak{d}(\mathsf{canWalk}) = 0.3$ and $\mathfrak{d}(\mathsf{canSail}) = 0.3$. Let $\mathsf{ActivePlace}$, $\mathsf{Mangrove}$, Beach, Place, $\mathsf{Trekking}$, $\mathsf{Kayaking}$, $\mathsf{canWalk}$, and $\mathsf{canSail}$ be rewritten shortly as AP, M, B, P, T, K, cW, and cS, respectively. Using Definition 13,

$$
\begin{aligned}
\mathsf{hd}^{\pi}(\mathcal{T}_{\mathsf{AP}}, \mathcal{T}_{\mathsf{M}}) &= \left(\frac{3}{6}\right) \cdot \mathsf{p\text{-}hd}^{\pi}(\mathcal{P}_{\mathsf{AP}}, \mathcal{P}_{\mathsf{M}}) + \left(\frac{3}{6}\right) \cdot \mathsf{e\text{-}set\text{-}hd}^{\pi}(\mathcal{E}_{\mathsf{AP}}, \mathcal{E}_{\mathsf{M}}) \\
&= \left(\frac{3}{6}\right) \cdot \left(\frac{\mathfrak{i}(X) \cdot \max\{\mathfrak{s}(X, Y), \mathfrak{s}(X, \mathsf{P})\} + \mathfrak{i}(\mathsf{P}) \cdot \max\{\mathfrak{s}(\mathsf{P}, Y), \mathfrak{s}(\mathsf{P}, \mathsf{P})\}}{\mathfrak{i}(X) + \mathfrak{i}(\mathsf{P})}\right) \\
&\quad + \left(\frac{3}{6}\right) \cdot \mathsf{e\text{-}set\text{-}hd}^{\pi}(\mathcal{E}_{\mathsf{AP}}, \mathcal{E}_{\mathsf{M}}) \\
&= \left(\frac{3}{6}\right) \left(\frac{1 \cdot \max\{0, 0\} + 2 \cdot \max\{0, 1\}}{1 + 2}\right) + \left(\frac{3}{6}\right) \cdot \mathsf{e\text{-}set\text{-}hd}^{\pi}(\mathcal{E}_{\mathsf{AP}}, \mathcal{E}_{\mathsf{M}}) \\
&= \left(\frac{3}{6}\right) \left(\frac{2}{3}\right) \\
&\quad + \left(\frac{3}{6}\right) \left[\frac{\mathfrak{i}(\mathsf{cW}) \cdot \max\{\mathsf{e\text{-}hd}^{\pi}(\exists \mathsf{cW}.\mathsf{T}, \exists \mathsf{cW}.\mathsf{T})\} + 1 \cdot \max\{0.019\}}{\mathfrak{i}(\mathsf{cW}) + \mathfrak{i}(\mathsf{cS})}\right] \\
&= \left(\frac{3}{6}\right) \left(\frac{2}{3}\right) + \left(\frac{3}{6}\right) \left[\frac{2 \cdot \max\{(1)(0.3 + 0.7(1))\} + 1 \cdot \max\{0.019\}}{\mathfrak{i}(\mathsf{cW}) + \mathfrak{i}(\mathsf{cS})}\right] \\
&= \left(\frac{3}{6}\right) \left(\frac{2}{3}\right) + \left(\frac{3}{6}\right) \left[\frac{(2)(1) + (1)(0.019)}{2 + 1}\right] \approx 0.67
\end{aligned}
$$

Similarly, we obtain $\mathsf{hd}^{\pi}(\mathcal{T}_{\mathsf{M}}, \mathcal{T}_{\mathsf{AP}}) = 0.80$. Hence, $\mathsf{sim}^{\pi}(\mathsf{M}, \mathsf{AP}) \approx 0.74$ by Definition 14. Furthermore, using Definition 13, $\mathsf{hd}^{\pi}(\mathcal{T}_{\mathsf{AP}}, \mathcal{T}_{\mathsf{B}}) \approx 0.51$ and $\mathsf{hd}^{\pi}(\mathcal{T}_{\mathsf{B}}, \mathcal{T}_{\mathsf{AP}}) = 0.75$. Hence, $\mathsf{sim}^{\pi}(\mathsf{B}, \mathsf{AP}) \approx 0.63$ by Definition 14.

The fact that $\mathsf{sim}^{\pi}(\mathsf{M}, \mathsf{AP}) > \mathsf{sim}^{\pi}(\mathsf{B}, \mathsf{AP})$ corresponds with the agent A's needs and preferences.

4.2 Properties of sim^{π}

Previously, we theorize a set of desirable properties that a concept similarity measure under preference profile should satisfy and formally introduce the measure sim^{π}. In this subsection, we provide mathematical proofs for the properties of sim^{π}. This gives many benefits to the users of sim^{π} since they can predict its expected behaviors.

Lemma 1. For $\mathcal{T}_D, \mathcal{T}_C \in \mathbf{T}^{\mathcal{ELH}}$, $\mathsf{hd}^{\pi_0}(\mathcal{T}_D, \mathcal{T}_C) = \mathsf{hd}(\mathcal{T}_D, \mathcal{T}_C)$.

Proof. (Sketch) Recall by Definition 10 that the default preference profile π_0 is the quintuple $\langle \mathfrak{i}_0^{\mathfrak{c}}, \mathfrak{i}_0^{\mathfrak{r}}, \mathfrak{s}_0^{\mathfrak{c}}, \mathfrak{s}_0^{\mathfrak{r}}, \mathfrak{d}_0 \rangle$. Also, suppose a concept name D is of the form: $P_1 \sqcap \cdots \sqcap P_m \sqcap \exists r_1.D_1 \sqcap \cdots \sqcap \exists r_n.D_n$, where $P_i \in \mathsf{CN}^{\mathsf{pri}}$, $r_j \in \mathsf{CN}$, $D_j \in \mathsf{Con}(\mathcal{ELH})$, $1 \leq i \leq m$, $1 \leq j \leq n$, $P_1 \sqcap \cdots \sqcap P_m$ is denoted by $\mathcal{P}_D$, and $\exists r_1.D_1 \sqcap \cdots \sqcap \exists r_n.D_n$ is denoted by $\mathcal{E}_D$. Let d be the depth of $\mathcal{T}_D$. We prove that, for any $d \in \mathbb{N}$, $\mathsf{hd}^{\pi_0}(\mathcal{T}_D, \mathcal{T}_C) = \mathsf{hd}(\mathcal{T}_D, \mathcal{T}_C)$ by induction on d.

When $d = 0$, we know that $D = P_1 \sqcap \cdots \sqcap P_m$. To show that $\mathsf{hd}^{\pi_0}(\mathcal{T}_D, \mathcal{T}_C) = \mathsf{hd}(\mathcal{T}_D, \mathcal{T}_C)$, we need to show that $\mu^{\pi_0} = \mu$ and $\mathsf{p\text{-}hd}^{\pi_0}(\mathcal{P}_D, \mathcal{P}_C) = \mathsf{p\text{-}hd}(\mathcal{P}_D, \mathcal{P}_C)$. Let us derive as follows:

$$\mu^{\pi_0} = \frac{\sum_{A \in \mathcal{P}_D} \hat{\mathfrak{i}}(A)}{\sum_{A \in \mathcal{P}_D} \hat{\mathfrak{i}}(A) + \sum_{\exists r.X \in \mathcal{E}_D} \hat{\mathfrak{i}}(r)} = \frac{\sum_{i=1}^{m} 1}{\sum_{i=1}^{m} 1 + 0} = \frac{m}{m+0} = \mu.$$

Furthermore, we only need to show $\sum_{A \in \mathcal{P}_D} \max\{\hat{\mathfrak{s}}(A, B) : B \in \mathcal{P}_C\} = |\mathcal{P}_D \cap \mathcal{P}_C|$ in order to show $\mathsf{p\text{-}hd}^{\pi_0}(\mathcal{P}_D, \mathcal{P}_C) = \mathsf{p\text{-}hd}(\mathcal{P}_D, \mathcal{P}_C)$. We know that $\mathfrak{s}_0^{\mathfrak{c}}$ maps name identity to 1 and otherwise to 0. Thus, $\sum_{A \in \mathcal{P}_D} \max\{\hat{\mathfrak{s}}(A, B) : B \in \mathcal{P}_C\} = |\{x : x \in \mathcal{P}_D \text{ and } x \in \mathcal{P}_C\}| = |\mathcal{P}_D \cap \mathcal{P}_C|$.

We must now prove that if $\mathsf{hd}^{\pi_0}(\mathcal{T}_D, \mathcal{T}_C) = \mathsf{hd}(\mathcal{T}_D, \mathcal{T}_C)$ holds for $d = h - 1$ where $h > 1$ and $D = P_1 \sqcap \cdots \sqcap P_m \sqcap \exists r_1.D_1 \sqcap \cdots \sqcap \exists r_n.D_n$ then $\mathsf{hd}^{\pi_0}(\mathcal{T}_D, \mathcal{T}_C) = \mathsf{hd}(\mathcal{T}_D, \mathcal{T}_C)$ also holds for $d = h$. To do that, we have to show $\mathsf{e\text{-}set\text{-}hd}^{\pi_0}(\mathcal{E}_D, \mathcal{E}_C) = \mathsf{e\text{-}set\text{-}hd}(\mathcal{E}_D, \mathcal{E}_C)$. This can be done by showing in the similar manner that $\gamma^{\pi_0} = \gamma$ and $\mathsf{hd}^{\pi_0}(\mathcal{T}_X, \mathcal{T}_Y) = \mathsf{hd}(\mathcal{T}_X, \mathcal{T}_Y)$ from $\mathsf{e\text{-}hd}^{\pi_0}(\exists r.X, \exists s.Y) = \mathsf{e\text{-}hd}(\exists r.X, \exists s.Y)$, where $\exists r.X \in \mathcal{E}_D$ and $\exists s.Y \in \mathcal{E}_C$. Consequently, it follows by induction that, for $\mathcal{T}_D, \mathcal{T}_C \in \mathbf{T}^{\mathcal{ELH}}$, $\mathsf{hd}^{\pi_0}(\mathcal{T}_D, \mathcal{T}_C) = \mathsf{hd}(\mathcal{T}_D, \mathcal{T}_C)$. □

Theorem 3. Let $C, D \in \mathsf{Con}(\mathcal{ELH})$, $\mathsf{sim}^{\pi_0}(C, D) = \mathsf{sim}(C, D)$.

Proof. It immediately follows from Lemma 1, Definition 4, and Definition 14. □

Theorem 3 tells us that sim^{π} is backward compatible in the sense that using sim^{π} with $\pi = \pi_0$, i.e. sim^{π_0}, coincides with sim. Technically speaking, sim^{π_0} can be used to handle the case of similar concepts regardless of the agent's preferences.

Theorem 4. sim^{π} is *symmetric*.

Proof. Let Π be a countably infinite set of preference profile. Fix any $\pi \in \Pi$ and $C, D \in \mathsf{Con}(\mathcal{ELH})$, we have $\mathsf{sim}^\pi(C, D) = \mathsf{sim}^\pi(D, C)$ by Definition 14. □

Theorem 5. sim^π is *equivalence invariant*.

Proof. Let Π be a countably infinite set of preference profile. Fix any $\pi \in \Pi$ and $C, D, E \in \mathsf{Con}(\mathcal{ELH})$, we show $C \equiv D \Longrightarrow \mathsf{sim}^\pi(C, E) = \mathsf{sim}^\pi(D, E)$.

Suppose $C \equiv D$, i.e. $C \sqsubseteq D$ and $D \sqsubseteq C$, then we know there exists a homomorphism $h_1 : \mathcal{T}_D \to \mathcal{T}_C$ which maps the root of $\mathcal{T}_D$ to the root of $\mathcal{T}_C$ and $h_2 : \mathcal{T}_C \to \mathcal{T}_D$ which maps the root of $\mathcal{T}_C$ to the root of $\mathcal{T}_D$, respectively, by Theorem 1. This means $\mathcal{T}_C = \mathcal{T}_D$. Thus, $\mathsf{sim}^\pi(C, E) = \mathsf{sim}^\pi(D, E)$. □

Theorem 6. sim^π is *structurally dependent*.

Proof. (Sketch) Let Π be a countably infinite set of preference profile. Fix any $\pi \in \Pi$ and any finite sets of concepts C_1 and C_2 with the following conditions:

1. $\mathsf{C}_1 \subseteq \mathsf{C}_2$;
2. concepts $A, B \notin \mathsf{C}_2$;
3. $\mathfrak{i}^\mathfrak{c}(\Phi) > 0$ if primitive $\Phi \in \mathsf{C}_2$;
4. $\mathfrak{i}^\mathfrak{r}(\varphi) > 0$ if existential $\exists\varphi.\Psi \in \mathsf{C}_2$.

Suppose $C \coloneqq \bigsqcap(\mathsf{C}_1 \cup \{A\})$, $D \coloneqq \bigsqcap(\mathsf{C}_1 \cup \{B\})$, $E \coloneqq \bigsqcap(\mathsf{C}_2 \cup \{A\})$ and $F \coloneqq \bigsqcap(\mathsf{C}_2 \cup \{B\})$ where $\mathsf{C}_1 = \{P_1, \ldots, P_m, \exists r_1.P'_1, \ldots, \exists r_n.P'_n\}$ and $\mathsf{C}_2 = \{P_1, \ldots, P_i, \exists r_1.P'_1, \ldots, \exists r_j.P'_j\}$, w.l.o.g. we show $\mathsf{sim}^\pi(C, D) \leq \mathsf{sim}^\pi(E, F)$ by following two cases.

Suppose $m \leq i$, $n = j$, and A, B be primitives, we have $\mathsf{p\text{-}hd}^\pi(\mathcal{P}_C, \mathcal{P}_D) = \frac{\sum_{P \in \mathcal{P}_C} \mathfrak{i}^\mathfrak{c}(P)}{\sum_{P \in \mathcal{P}_C} \mathfrak{i}^\mathfrak{c}(P) + \mathfrak{i}^\mathfrak{c}(A)}$, $\mathsf{p\text{-}hd}^\pi(\mathcal{P}_D, \mathcal{P}_C) = \frac{\sum_{P \in \mathcal{P}_D} \mathfrak{i}^\mathfrak{c}(P)}{\sum_{P \in \mathcal{P}_D} \mathfrak{i}^\mathfrak{c}(P) + \mathfrak{i}^\mathfrak{c}(B)}$, $\mathsf{p\text{-}hd}^\pi(\mathcal{P}_E, \mathcal{P}_F) = \frac{\sum_{P \in \mathcal{P}_E} \mathfrak{i}^\mathfrak{c}(P)}{\sum_{P \in \mathcal{P}_E} \mathfrak{i}^\mathfrak{c}(P) + \mathfrak{i}^\mathfrak{c}(A)}$, and $\mathsf{p\text{-}hd}^\pi(\mathcal{P}_F, \mathcal{P}_E) = \frac{\sum_{P \in \mathcal{P}_F} \mathfrak{i}^\mathfrak{c}(P)}{\sum_{P \in \mathcal{P}_F} \mathfrak{i}^\mathfrak{c}(P) + \mathfrak{i}^\mathfrak{c}(B)}$.

Since $m \leq i$, we know $\mathsf{p\text{-}hd}^\pi(\mathcal{P}_C, \mathcal{P}_D) \leq \mathsf{p\text{-}hd}^\pi(\mathcal{P}_E, \mathcal{P}_F)$ and $\mathsf{p\text{-}hd}^\pi(\mathcal{P}_D, \mathcal{P}_C) \leq \mathsf{p\text{-}hd}^\pi(\mathcal{P}_F, \mathcal{P}_E)$. This infers $\mathsf{sim}^\pi(C, D) \leq \mathsf{sim}^\pi(E, F)$.

Suppose $m = i$, $n \leq j$, and A, B be existentials, then with the similar manner, we can show $\mathsf{e\text{-}set\text{-}hd}^\pi(\mathcal{E}_C, \mathcal{E}_D) \leq \mathsf{e\text{-}set\text{-}hd}^\pi(\mathcal{E}_E, \mathcal{E}_F)$ and $\mathsf{e\text{-}set\text{-}hd}^\pi(\mathcal{E}_D, \mathcal{E}_C) \leq \mathsf{e\text{-}set\text{-}hd}^\pi(\mathcal{E}_F, \mathcal{E}_E)$. This also infers $\mathsf{sim}^\pi(C, D) \leq \mathsf{sim}^\pi(E, F)$.

Therefore, we have shown $\mathsf{sim}^\pi(C, D) \leq \mathsf{sim}^\pi(E, F)$. □

Lemma 2. Let $\mathcal{T}_D, \mathcal{T}_C \in \mathbf{T}^{\mathcal{ELH}}$ and Π be a countably infinite set of preference profile. Then, $\mathsf{hd}(\mathcal{T}_D, \mathcal{T}_C) = 1 \Longleftrightarrow \forall \pi \in \Pi : \mathsf{hd}^\pi(\mathcal{T}_D, \mathcal{T}_C) = 1$.

Proof. Let Π be a countably infinite set of preference profile and π_0 be the default preference profile. Fix any $\pi \in \Pi$, we show $\mathsf{hd}(\mathcal{T}_D, \mathcal{T}_C) = 1 \Longleftrightarrow \mathsf{hd}^\pi(\mathcal{T}_D, \mathcal{T}_C) = 1$.

($\Rightarrow$) $\mathsf{hd}(\mathcal{T}_D, \mathcal{T}_C) = 1$ implies that there exists a homomorphism $h : \mathcal{T}_D \to \mathcal{T}_C$ which maps the root of $\mathcal{T}_D$ to the root of $\mathcal{T}_C$. Consequently, any setting on π does not influence the calculation on $\mathsf{hd}^\pi(\mathcal{T}_D, \mathcal{T}_C)$.

($\Leftarrow$) In particular, it suffices to show $\mathsf{hd}^{\pi_0}(\mathcal{T}_D, \mathcal{T}_C) = 1 \Longrightarrow \mathsf{hd}(\mathcal{T}_D, \mathcal{T}_C) = 1$. By Lemma 1, it is the case that $\mathsf{hd}(\mathcal{T}_D, \mathcal{T}_C) = 1$.

□

Theorem 7. sim^{π} is *preference invariant w.r.t. equivalence.*

Proof. (Sketch) Let $C, D \in \mathsf{Con}(\mathcal{ELH})$ and Π be a countably infinite set of preference profile. Fix any $\pi \in \Pi$, we show $C \equiv D \Longleftrightarrow \mathsf{sim}^{\pi}(C, D) = 1$.

($\Rightarrow$) Assume $C \equiv D$, we need to show $\mathsf{sim}^{\pi}(C, D) = 1$. By Theorem 2, we know $C \equiv D \Longleftrightarrow \mathsf{sim}(C, D) = 1$. With the usage of Lemma 2, Definition 4, and Definition 14, we can derive $\mathsf{sim}^{\pi}(C, D) = 1$.

($\Leftarrow$) This can be shown similarly as in the forward direction.

□

Theorem 4 to 7 spells out that sim^{π} satisfies all fundamental properties of concept similarity measure under preference profile.

Another important property of sim^{π} is that there exists an algorithmic procedure whose execution time is upper bounded by a polynomial expression in the size of the description trees (Theorem 8).

Theorem 8. Assume that a value from any preference functions is retrieved in $\mathcal{O}(1)$. Given $C, D \in \mathsf{Con}(\mathcal{ELH})$, $\mathsf{sim}^{\pi}(C, D) \in \mathcal{O}(|V_C| \cdot |V_D|)$ where V_C and V_D are set of vertices of the description trees $\mathcal{T}_C$ and $\mathcal{T}_D$, respectively.

Proof. (Sketch) Let $C, D \in \mathsf{Con}(\mathcal{ELH})$, π be any preference profile, and $\mathcal{T}_C, \mathcal{T}_D$ be corresponding description trees. By Definition 14, we show $\mathsf{hd}^{\pi}(\mathcal{T}_C, \mathcal{T}_D) \in \mathcal{O}(|V_C| \cdot |V_D|)$ and $\mathsf{hd}^{\pi}(\mathcal{T}_D, \mathcal{T}_C) \in \mathcal{O}(|V_D| \cdot |V_C|)$. W.l.o.g. it suffices to show merely $\mathsf{hd}^{\pi}(\mathcal{T}_C, \mathcal{T}_D) \in \mathcal{O}(|V_C| \cdot |V_D|)$, i.e., we show the computation of each composing part is upper bounded by $|V_C| \cdot |V_D|$. □

Definition 12 suggests that different preference profile settings represent different types of a rational agent. An easy characterization is observed from the aspect of role discount factor ($\mathfrak{d}$). Intuitively, when the settings $\mathfrak{i}^{\mathfrak{c}}$, $\mathfrak{i}^{\mathfrak{r}}$, $\mathfrak{s}^{\mathfrak{c}}$, and $\mathfrak{s}^{\mathfrak{r}}$ defined by two rational agents A, B are the same, the agent which defines the lower $\mathfrak{d}$ on every $r \in \mathsf{RN}$ is always more skeptical. For instance, if $\mathfrak{d}_{\mathrm{A}}(\mathsf{canWalk}) = 0.3$ and $\mathfrak{d}_{\mathrm{B}}(\mathsf{canWalk}) = 0.4$, then $\mathsf{sim}^{\pi_{\mathrm{A}}}(\exists \mathsf{canWalk}.\mathsf{Trekking}, \exists \mathsf{canWalk}.\mathsf{Parading}) = 0.3$ and $\mathsf{sim}^{\pi_{\mathrm{B}}}(\exists \mathsf{canWalk}.\mathsf{Trekking}, \exists \mathsf{canWalk}.\mathsf{Parading}) = 0.4$. This is clear that the agent A is more skeptical than the agent B.

Proposition 1. Let Π be a countably infinite set of preference profile and $\pi_1, \pi_2 \in \Pi$ such that $\pi_1 = \langle \mathfrak{i}_1^{\mathfrak{c}}, \mathfrak{i}_1^{\mathfrak{r}}, \mathfrak{s}_1^{\mathfrak{c}}, \mathfrak{s}_1^{\mathfrak{r}}, \mathfrak{d}_1 \rangle$, $\pi_2 = \langle \mathfrak{i}_2^{\mathfrak{c}}, \mathfrak{i}_2^{\mathfrak{r}}, \mathfrak{s}_2^{\mathfrak{c}}, \mathfrak{s}_2^{\mathfrak{r}}, \mathfrak{d}_2 \rangle$, and RN be a set of role names. The following holds:

$$\forall r \in \mathsf{RN} : (\mathfrak{d}_1(r) \leq \mathfrak{d}_2(r)) \Longrightarrow \equiv \preceq \ \mathsf{sim}^{\pi_1} \ \preceq \ \mathsf{sim}^{\pi_2}$$

for fixed functions $\mathfrak{i}_1^{\mathfrak{c}} = \mathfrak{i}_2^{\mathfrak{c}}$, $\mathfrak{i}_1^{\mathfrak{r}} = \mathfrak{i}_2^{\mathfrak{r}}$, $\mathfrak{s}_1^{\mathfrak{c}} = \mathfrak{s}_2^{\mathfrak{c}}$, and $\mathfrak{s}_1^{\mathfrak{r}} = \mathfrak{s}_2^{\mathfrak{r}}$.

5 IMPLEMENTATION METHODS OF SIM^{π}

Theorem 8 tells us that sim^{π} can be computed in the polynomial time. This section exhibits two algorithmic procedures of sim^{π} belonging to that class.

5.1 Top-Down Implementation of sim^{π}

Algorithm 1 Pseudo code for hd^{π} using top-down fashion

function $\mathsf{hd}^{\pi}(\mathcal{T}_D, \mathcal{T}_C, \pi)$
  return $(\mu^{\pi}(\mathcal{T}_D, \pi) \times \mathsf{p\text{-}hd}^{\pi}(\mathcal{P}_D, \mathcal{P}_C, \pi)) + ((1 - \mu^{\pi}(\mathcal{T}_D, \pi)) \times \mathsf{e\text{-}set\text{-}hd}^{\pi}(\mathcal{E}_D, \mathcal{E}_C, \pi))$
end function

function $\mathsf{e\text{-}set\text{-}hd}^{\pi}(\mathcal{E}_D, \mathcal{E}_C, \pi)$
  if $\sum \mathfrak{i}^{\mathfrak{r}}(\mathcal{E}_D, \pi) = 0$ **then**
    return 1
  else if $\sum \mathfrak{i}^{\mathfrak{r}}(\mathcal{E}_C, \pi) = 0$ **then**
    return 0
  else
    $w \leftarrow 0$
    for $\exists r.X \in \mathcal{E}_D$ **do**
      $m \leftarrow 0$
      for $\exists s.Y \in \mathcal{E}_C$ **do**
        $e \leftarrow \mathsf{e\text{-}hd}^{\pi}(\exists r.X, \exists s.Y, \pi)$
        if $e > m$ **then**
          $m \leftarrow e$
        end if
      end for
      $w \leftarrow w + (m \times \hat{\mathfrak{i}}(r))$
    end for
    return $w / \sum \mathfrak{i}^{\mathfrak{r}}(\mathcal{P}_D, \pi)$
  end if
end function

function $\mathsf{e\text{-}hd}^{\pi}(\exists r.X, \exists s.Y, \pi)$
  return $\gamma^{\pi}(r, s, \pi) \times (\hat{\mathfrak{d}}(r) + ((1 - \hat{\mathfrak{d}}(r)) \times \mathsf{hd}^{\pi}(\mathcal{T}_X, \mathcal{T}_Y, \pi)))$
end function

Algorithm 1 presents the top-down approach for sim^{π} implementation. Due to the limited space, we omit to show Algorithm 1 in details. The reader may easily observe that the time efficiency of Algorithm 1 is quintic because the computation of $\mathsf{p\text{-}hd}^{\pi}$ is quadratic and $\mathsf{e\text{-}set\text{-}hd}^{\pi}$ contains double nested loops which indirectly

make recursive calls to hd^{π}. It is also not difficult to observe that the number of recursive calls is upper bounded by the height of the description trees.

It is worth to mention that using hd^{π} requires concept descriptions to be transformed into $\mathcal{ELH}$ description trees. Taking this as an advantage, the next subsection introduces an alternative way to compute hd^{π} from bottom to up, which is approximately three times faster than the counterpart top-down approach in the worst case (cf. Subsection 6.1 for useful discussion).

5.2 Bottom-Up Implementation of sim^{π}

Rather than computing (possibly duplicated) value of hd^{π} again and again, Algorithm 2 employs the classical bottom-up version of dynamic programming technique to compute hd^{π} of the smaller subtrees and records the results in a table (see the variable *result*[·][·] in Algorithm 2) from which a solution to the original computation of hd^{π} can be then obtained (cf. at line No. 20, the function returns value *result*[0][0]).

To compute hd^{π} from bottom to up, we need to know the height of the trees in advance. For Algorithm 2, we employ *breath-first search* algorithm (denoted by BFS) to determine the height of each description tree (cf. line No. 4 and 5 of the algorithm). Algorithm 2 reuses the methods μ^{π}, $\mathsf{p\text{-}hd}^{\pi}$, $\mathsf{e\text{-}set\text{-}hd}^{\pi}$, γ^{π}, $\sum \mathfrak{i}^{\mathfrak{c}}$, and $\sum \mathfrak{i}^{\mathfrak{r}}$ from Algorithm 1 and provides pseudo code for $\mathsf{e\text{-}hd}^{\pi}$ since it is merely overridden.

What is the time complexity of Algorithm 2? It should be quintic because the algorithm considers the similarity of all the different pairs of two concept names for h times (cf. line No. 6). More formally, we know $\mathit{result}[\mathcal{T}_{\gamma}][\mathcal{T}_{\lambda}] \in \mathcal{O}(v^2)$ where v denotes the set cardinality of $\mathcal{P}_x$ (and $\mathcal{E}_x$) for any description tree x. Let $m(i)$ and $n(i)$ be the number of nodes on level i of description trees D and C, respectively. Then, the number of times operation *result*[·][·] is executed (say C) is equal to:

$$
\begin{aligned}
C &= \sum_{i=0}^{h-1}\sum_{j=0}^{m(i)}\sum_{k=0}^{n(i)} v^2 \\
&= v^2 \sum_{i=0}^{h-1}\sum_{j=0}^{m(i)}\sum_{k=0}^{n(i)} 1 \\
&= v^2 \sum_{i=0}^{h-1}\sum_{j=0}^{m(i)} (n(i)+1) \\
&= v^2 \sum_{i=0}^{h-1} (n(i)+1)(m(i)+1) \\
&= v^2 \big[[(n(0)+1)(m(0)+1)] + [(n(1)+1)(m(1)+1)] \\
&\quad + \ldots + [(n(h-1)+1)(m(h-1)+1)]\big].
\end{aligned}
$$

Algorithm 2 Pseudo code for hd^{π} using bottom-up fashion

```
1: Initialize a global result[·][·] to store the degree of similarity between 2 concepts.
2:
3: function hd^π(T_D, T_C, π)
4:     Map < ℤ, List < T >> map_D ← BFS(T_D)        ▷ map_D stores nodes on each level of T_D
5:     Map < ℤ, List < T >> map_C ← BFS(T_C)        ▷ map_C stores nodes on each level of T_C
6:     h ← map_D.size()
7:     for i = h − 1 to 0 do
8:         List < T > list_Tγ ← map_D.get(i)
9:         List < T > list_Tλ ← map_C.get(i)
10:        for T_γ ∈ list_Tγ do
11:            for list_Tλ ≠ null and T_λ ∈ list_Tλ do
12:                if i = h − 1 then
13:                    result[T_γ][T_λ] ← p-hd^π(P_γ, P_λ, π)
14:                else
15:                    result[T_γ][T_λ] ← (μ^π(T_γ, π) × p-hd^π(P_γ, P_λ, π))
       + ((1 − μ^π(T_γ, π)) × e-set-hd^π(E_γ, E_λ, π))
16:                end if
17:            end for
18:        end for
19:    end for
20:    return result[0][0]
21: end function
22:
23: function e-hd^π(∃r.X, ∃s.Y, π)
24:    hd′ ← result[T_X][T_Y]
25:    if hd′ = null then
26:        hd′ ← 0
27:    end if
28:    return γ^π(r, s, π) × (δ̂(r) + ((1 − δ̂(r)) × hd′))
29: end function
```

Thus, the algorithm makes the similar number of operations as Algorithm 1, plus an additional amount of extra space. On the positive side, the algorithm has never recursively invoked itself to determine the similarity of different pairs of nested concepts, i.e., it directly uses values stored in the table. The algorithm also shows that computing the similarity of nodes from level i, where i is greater than the minimum height of description trees (cf. the condition $list_{\mathcal{T}_\Lambda}! = \text{null}$ at line No. 11), is irrelevant to the computation.

Algorithm 2 does work productively in an environment where recursion is fairly expensive. For example, imperative languages, such as Java, C, and Python, are

typically faster if using a loop and slower if doing a recursion. On the other hand, for some implementations of functional programming languages, iterations may be very expensive and recursion may be very cheap. In many implementations of them, recursion is transformed into a simple jump but changing the loop variables (which are mutable) requires heavy operations. Subsection 6.1 reports that the practical performance agrees to this theoretical analysis that the bottom-up approach is more efficient when implemented by imperative languages, such as Java.

6 EMPIRICAL EVALUATION

This section evaluates the practical performance of both algorithms against sim^5, reassures pragmatically the backward compatibility of sim^π under π_0 (Theorem 3 already proves this), and discusses the applicability of sim^π in potential use cases.

6.1 Performance Analysis and Backward Compatibility of sim^π

Both versions of sim^π (cf. Subsection 5.1 and Subsection 5.2) are implemented in Java version 1.8 with the usage of Spring Boot version 1.3.3.RELEASE. All the dependencies are managed by Apache Maven version 3.2.5. We also implement unit test cases along with the development of both versions to verify the correctness of their behaviors. In the current state (when we are writing this paper), there are 111 unit test cases. All of them are written to cover important parts of both implementations.

To perform benchmarking, we have selected SNOMED CT as a test ontology. As mentioned in the introduction, it is one of the largest and the most widely used medical ontologies currently available, and also, is expressible in $\mathcal{ELH}$. In our experiments, we employ a SNOMED CT ontology version from January 2005 (hitherto referred as $\mathcal{O}_{\text{SNOMED}}$) which contains 379 691 concept names and 62 role names. Moreover, each defined concept is categorized into the 18 mutually exclusive top-level concepts. In the sense of subsumption relation, concepts belonging to the same category should be more similar than those belonging to different categories.

For our experiments, we used a 2.4 GHz Intel Core i5 with 8 GB RAM under OS X El Capitan. Unfortunately, the overall number of concept pairs in $\mathcal{O}_{\text{SNOMED}}$ is approximately 10^{11}. Suppose an execution of sim^π takes around a millisecond, we still need around 1 158 days in order to complete the entire ontology. According to this reason, we consider 2 out of 18 categories, viz. *Clinical Finding* and *Procedure*, although there are more category pairs. Then, we randomly select 0.5 % of *Clinical Finding*, i.e. 206 concepts, denoted by $\mathbf{C}'_1$. After that, we randomly select the same number of concepts from *Procedure*, i.e. 206 concepts, denoted by $\mathbf{C}'_2$. This sampled set is denoted by $\mathcal{O}'_{\text{SNOMED}}$, i.e. $\mathcal{O}'_{\text{SNOMED}} = \mathbf{C}'_1 \cup \mathbf{C}'_2$. Then, we create three test datasets from this sampled set, viz. $\mathbf{C}'_1 \times \mathbf{C}'_1$, $\mathbf{C}'_1 \times \mathbf{C}'_2$, and $\mathbf{C}'_2 \times \mathbf{C}'_2$.

[5] We have re-implemented sim (proposed in [10]) based on the same technologies and techniques as sim^π.

Firstly, we estimate the practical performance of the top-down fashion. For each concept pair in each set, we

1. employ the default preference profile π_0 on (top-down) sim^{π};
2. measure the similarity of concepts in $\mathcal{O}'_{\text{SNOMED}}$ by peeking on $\mathcal{O}_{\text{SNOMED}}$ to help unfolding;
3. repeat the previous step with (top-down) sim;
4. repeat steps 2.–3. three times and calculate the statistical results (in milliseconds).

Results are gathered in Table 1. We note that *avg*, *max*, and *min* represent the execution time for measuring similarity of a concept pair in the average case, in the worst case, and in the best case, respectively.

Pairs	Number of Pairs	sim (avg/max/min)	sim^{π_0} (avg/max/min)
$\mathbf{C}'_1 \times \mathbf{C}'_1$	25	2.280/7.000/0.000	1.800/10.000/0.000
$\mathbf{C}'_1 \times \mathbf{C}'_2$	215	2.291/97.000/0.000	2.278/84.000/0.000
$\mathbf{C}'_2 \times \mathbf{C}'_2$	1 849	3.395/45.000/0.000	3.931/128.000/0.000

Table 1. Execution time of top-down sim and top-down sim^{π_0} on $\mathcal{O}'_{\text{SNOMED}}$

Secondly, we estimate the practical performance of the bottom-up fashion by following the same steps as we did previously. Indeed, we exclude the time used to determine the height of each description tree, i.e., our benchmark begins from line No. 7 to 21 of Algorithm 2. Table 2 gathers up the results.

Pairs	Number of Pairs	sim (avg/max/min)	sim^{π_0} (avg/max/min)
$\mathbf{C}'_1 \times \mathbf{C}'_1$	25	2.200/6.000/0.000	1.693/5.000/0.000
$\mathbf{C}'_1 \times \mathbf{C}'_2$	215	2.040/32.000/0.000	1.946/10.000/0.000
$\mathbf{C}'_2 \times \mathbf{C}'_2$	1 849	3.368/55.000/0.000	3.435/45.000/0.000

Table 2. Execution time of bottom-up sim and bottom-up sim^{π_0} on $\mathcal{O}'_{\text{SNOMED}}$

The experiment shows that the practical performance of sim^{π} is likely equal to the performance obtained by sim – as ones may not expect. The results show that the bottom-up sim^{π} performs approximately three times faster than the counterpart top-down sim^{π} (in the worst case) when implemented by imperative languages (e.g. Java as in our case). This conforms to our analysis discussed in Subsection 5.2.

Lastly, we evaluate the backward compatibility of sim^{π} with sim. Our goal is to ascertain that sim^{π} can be used interchangeably as the original sim by setting preference profile to the default one (Theorem 3 already proves this). To this point, we have performed an experiment on concept pairs defined in $\mathcal{O}'_{\text{SNOMED}}$. The experiment evaluates results from sim and sim^{π_0} and found that both coincide, as desired.

6.2 Applicability of sim^{π}

6.2.1 Tuning via $\mathfrak{i}^{\mathfrak{c}}$ and $\mathfrak{d}$

We show the applicability of $\mathfrak{i}^{\mathfrak{c}}$ and $\mathfrak{d}$ through similarity measuring on SNOMED CT. Figure 2 depicts an example unfoldable terminology extracted from $\mathcal{O}_{\text{SNOMED}}$.

$$
\begin{aligned}
\mathsf{NeonatalAspirationOfAmnioticFluid} &\equiv \mathsf{NeonatalAspirationSyndromes} \\
&\quad \sqcap \exists \mathsf{roleGroup}.(\exists \mathsf{causativeAgent}.\mathsf{AmnioticFluid}) \\
\mathsf{NeonatalAspirationOfMucus} &\equiv \mathsf{NeonatalAspirationSyndromes} \\
&\quad \sqcap \exists \mathsf{roleGroup}.(\exists \mathsf{causativeAgent}.\mathsf{Mucus}) \\
\mathsf{Hypoxemia} &\equiv \mathsf{DisorderOfRespiratorySystem} \sqcap \mathsf{DisorderOfBloodGas} \\
&\quad \sqcap \exists \mathsf{roleGroup}.(\exists \mathsf{interprets}.\mathsf{OxygenDelivery}) \\
&\quad \sqcap \exists \mathsf{roleGroup}.(\exists \mathsf{findingSite}.\mathsf{ArterialSystemStructure}) \\
\mathsf{BodySecretion} &\sqsubseteq \mathsf{BodySubstance} \\
\mathsf{BodySubstance} &\sqsubseteq \mathsf{Substance} \\
\mathsf{BodyFluid} &\sqsubseteq \mathsf{BodySubstance} \sqcap \mathsf{LiquidSubstance} \\
\mathsf{AmnioticFluid} &\sqsubseteq \mathsf{BodyFluid} \\
\mathsf{Mucus} &\sqsubseteq \mathsf{BodySecretion} \\
\mathsf{causativeAgent} &\sqsubseteq \mathsf{associatedWith}
\end{aligned}
$$

Figure 2. Example of $\mathcal{ELH}$ concept definitions defined in $\mathcal{O}_{\text{SNOMED}}$

Considering merely objective factors regardless of the agent's preferences, it yields that $\mathsf{sim}^{\pi_0}(\mathsf{NAOAF}, \mathsf{NAOM}) \approx 0.9$[6] and $\mathsf{sim}^{\pi_0}(\mathsf{NAOAF}, \mathsf{H}) = 0.2$. The results yield the quite similar concepts NAOAF and NAOM, which reflect the fact that both are resided in the same cluster of SNOMED CT. However, the result yielding that the concepts NAOAF and H share a little similarity controverts the fact that both carry neither implicit nor explicit relationship. This is indeed caused by the usage of the special-purpose role called roleGroup – informally read as *relation group*.

In SNOMED CT, the use of relation group is widely accepted to nestedly represent a group of existential information [21]. As a consequence, it increases unintentionally the degree of similarity due to role commonality (i.e. γ^{π}). Since roleGroup precedes every existential restriction, it is useless to regard an occurrence of this as being similar. The importance contribution of roleGroup in $\mathcal{O}_{\text{SNOMED}}$ should be none. Hence, the agent S who measures similarity on SNOMED CT should set $\mathfrak{d}_S(\mathsf{roleGroup}) = 0$.

Furthermore, the SNOMED CT top concept SCT-TOP subsumes every defined concept of each category. This means this special concept is shared by every expanded concept description. Intuitively, this special top concept is of no importance

[6] Obvious abbreviations are used here for the sake of succinctness.

for measuring similarity on SNOMED CT and we can treat the top-level concepts as directly subsumed by $\top$. As a result, the agent S should also set $\mathfrak{i}^{\mathfrak{c}}_S(\mathsf{SCT\text{-}TOP}) = 0$.

Tuning the measure with this expertise knowledge yields more realistic result. That is, the similarity of concepts under the same category which uses roleGroup in their definitions is slightly reduced. Also, the similarity of concepts under different categories is totally dissimilar. Continuing the case, $\mathsf{sim}^{\pi_S}(\mathsf{NAOAF}, \mathsf{NAOM}) \approx 0.84$ and $\mathsf{sim}^{\pi_S}(\mathsf{NAOAF}, \mathsf{H}) = 0.0$, as desired.

6.2.2 Tuning via $\mathfrak{s}^{\mathfrak{r}}$

Let us use the ontology given below to query for places similar to ActivePlace.

$$\mathsf{ActivePlace} \sqsubseteq \mathsf{Place} \sqcap \exists \mathsf{canSail.Kayaking}$$

$$\mathsf{Mangrove} \sqsubseteq \mathsf{Place} \sqcap \exists \mathsf{canWalk.Trekking}$$

$$\mathsf{Supermarket} \sqsubseteq \mathsf{Place} \sqcap \exists \mathsf{canBuy.FreshFood}$$

Suppose the agent feels *walking* and *sailing* are similar and are *still satisfied much* on both actions. Taking $\mathfrak{s}^{\mathfrak{r}}(\mathsf{canWalk}, \mathsf{canSail}) = 0.6$ yields $\mathsf{sim}^{\pi}(\mathsf{M}, \mathsf{AP}) > \mathsf{sim}^{\pi}(\mathsf{S}, \mathsf{AP})$, which conforms to the agent's preferences and needs.

6.2.3 Tuning via $\mathfrak{s}^{\mathfrak{c}}$

Let us use the ontology given below to query for a product which offers features the agent is satisfied with most.

$$\mathsf{WantedFeatures} \sqsubseteq \mathsf{F}_0 \sqcap \mathsf{F}_1 \sqcap \mathsf{F}_2$$

$$\mathsf{Item}_1 \sqsubseteq \mathsf{F}_0 \sqcap \mathsf{F}_3$$

$$\mathsf{Item}_2 \sqsubseteq \mathsf{F}_0 \sqcap \mathsf{F}_4$$

According to the ontology, WantedFeatures represents a collection of desired features and F_i (where $i \in \mathbb{N}$) represents a feature. A purchase decision is sometimes affected by satisfied alternations, which are varied by different people. Assume that the agent feels satisfaction to have F_3 if the agent cannot have F_1. Taking $\mathfrak{s}^{\mathfrak{c}}(\mathsf{F}_1, \mathsf{F}_3) = 0.8$ yields $\mathsf{sim}^{\pi}(\mathsf{WF}, \mathsf{I1}) > \mathsf{sim}^{\pi}(\mathsf{WF}, \mathsf{I2})$, which conforms to the agent's perceptions.

6.2.4 Tuning via $\mathfrak{i}^{\mathfrak{r}}$

Let us use the ontology given in Example 1 to query for places which are most similar to ActivePlace. Typically, a human decision is affected by a priority of concerns, which are varied by different people. Suppose that the agent weights more on places which permit to *walk* more than other activities. Taking $\mathfrak{i}^{\mathfrak{r}}(\mathsf{canWalk}) = 2$ yields $\mathsf{sim}^{\pi}(\mathsf{M}, \mathsf{AP}) > \mathsf{sim}^{\pi}(\mathsf{B}, \mathsf{AP})$, which conforms to the agent's preferences.

7 RELATED WORK

As we develop the notion $\overset{\pi}{\sim}_{\mathcal{T}}$ as a generalization of $\sim_{\mathcal{T}}$, this section relates our development to others in two areas, viz. CSMs without regard to the agent's preferences and CSMs with regard to the agent's preferences.

7.1 CSMs Without Regard to The Agent's Preferences

In the standard perception, CSM refers to the study of similar concepts inherited by nature, i.e. the ones similar regardless of the agent's preferences. CSM is widely studied and the techniques are roughly classified into two main groups, viz. path-distance-based approach and DLs-based approach.

In the path-distance-based approach, a degree of similarity is calculated based on the depth of a subsumption hierarchy. The method [22, 23] considers the distance between concepts w.r.t. their least common subsumer. A potential drawback of this approach is its ignorance on concept definitions defined in TBox. Hence, any pair of concepts out of the subsumption relation is always considered as totally dissimilar.

In DLs-based approach, a simple approach is developed in [20] for the DL $\mathcal{L}_0$ (i.e. no use of roles) and is known as Jaccard Index. Its extension to the DL $\mathcal{ELH}$ is proposed in [16]. This work also introduces important properties of CSM and suggests a general framework called *simi* which satisfied most of the properties. In *simi*, functions and operators, such as t-conorm and the fuzzy connector, are to be parameterized and thus left to be specified. The framework also does not contain implementation details. This may cause implementation difficulties since merely promising properties are given and no guideline of how concrete operators are chosen is provided. Similar approaches can be found in [4, 5, 6, 7] for other DLs.

There is another approach which considers their canonical interpretation of concepts in question, such as [8, 9]. A potential drawback of these approaches is that it cannot be applied to an ontology without ABox, e.g. SNOMED CT.

The notion of homomorphism degree is originally introduced in [13] and is thereof extended toward the development of sim^{π} in this work. Theorem 3 suggests that sim^{π} can be used to measure similarity of concepts inherently by nature through the setting π_0, i.e. sim^{π_0}. As inspired by the tree homomorphism, the measure differs [16] from the use of μ^{π} to determine how important the primitive concepts are to be considered and the use of γ^{π} to determine a degree of role commonality between matching edges of the description trees.

7.2 CSMs with Regard to The Agent's Preferences

Most CSMs are objective-based. However, there exists work [10, 16] which provides methodologies for tuning. We discuss their differences to ours in the following.

In an extended work of sim [10], a range of number for discount factor (ν) and the neglect of special concept names are used in the similarity application of SNOMED

CT. For instance, when roleGroup is found, the value of ν is set to 0. These ad hoc approaches can be viewed as specific applications of $\mathfrak{d}$ and $\mathfrak{i}^{\mathfrak{c}}$, respectively, of preference profile. Unfortunately, no other aspects of π appear in its use.

In *simi* [16], the function *pm* is used to define the similarity degree of primitive concept pairs and role pairs. Using *pm* with primitive concept pairs invokes the equivalent intuition as $\mathfrak{s}^{\mathfrak{c}}$; however, this does not mean so in the aspect $\mathfrak{s}^{\mathfrak{r}}$. Allowing to define the similarity of defined role names, as in [16], may be not appropriate since defined role names are contributed by primitive role names. For example, let $r_1 \sqsubseteq s_1$ and $r_2 \sqsubseteq s_2$ are defined in $\mathcal{T}$. It is clear that $r_1, r_2 \in \mathsf{RN}^{\mathsf{def}}$. By defining $pm(r_1, r_2)$, the defined similarity should be also propagated to the similarity of s_1 and s_2. However, this point is not discussed in [16]. In respect of this, $\mathsf{RN}^{\mathsf{pri}}$ is merely used in $\mathfrak{s}^{\mathfrak{r}}$ and γ^{π} is defined for the similarity of defined role names. The authors of [16] also define the function $g : N_A \to \mathbb{R}_{>0}$ representing the weight for concept names and existential restriction atoms (based on their definition). Ones may feel the resemblance of g and $\mathfrak{i}^{\mathfrak{c}}, \mathfrak{i}^{\mathfrak{r}}$; however, they are also different in three perspectives. Firstly, the mapping of g is reached to the infinity whereas $\mathfrak{i}^{\mathfrak{c}}$ and $\mathfrak{i}^{\mathfrak{r}}$ are bounded. This characteristic of g is impractical to use as it may lead to the unbalance of weight assignments. For instance, one may define $g(A_1) = 1$ but $g(A_2) = 10^{12}$ where $A_1, A_2 \in \mathsf{CN}^{\mathsf{pri}}$. To avoid this situation, the authors should provide a guideline for weight assignments. Secondly, the mapping of g is lower bounded by one. This clearly makes an impossibility to define the intuition of having no importance. Thus, the situation given in Subsubsection 6.2.4 is not expressible. Lastly, the domain of g is the set of atoms whereas $\mathfrak{i}^{\mathfrak{c}}$ (and $\mathfrak{i}^{\mathfrak{r}}$) is the set of primitive concept names (and the set of role names, respectively). Using the set of atoms as the domain is also impractical since there can be infinitely many existential restriction atoms and the interpretation of functions is slightly dubious. For instance, given $g(\exists r.C) = 2$ and $g(\exists r.D) = 3$, do both r intentionally contribute the equal importance? Thus, this definition is inappropriate to represent the agent's perception. Moreover, the aspect $\mathfrak{d}$ disappears from [16]. Lacking of fully $\mathfrak{i}^{\mathfrak{c}}$ and $\mathfrak{d}$ makes the framework inappropriate to use for SNOMED CT applications. These distinctions of *simi* and sim^{π} are radically caused by their different motivations. Table 3 summarizes this discussion, where ✔ denotes totally identical to the specified function whereas ✓ denotes partially identical to the specified function.

CSM	$\mathfrak{i}^{\mathfrak{c}}$	$\mathfrak{i}^{\mathfrak{r}}$	$\mathfrak{s}^{\mathfrak{c}}$	$\mathfrak{s}^{\mathfrak{r}}$	$\mathfrak{d}$
sim^{π}	✔	✔	✔	✔	✔
the extended work of sim [10]	✔				✔
simi [16]	✓		✔	✓	

Table 3. Concept similarity measures which embed preference elements

Not only distinct on the mathematical representation of *simi* and sim^{π}, the desired properties presented in each work are also different. While the properties introduced in [16] are motivated for CSM, our properties are developed under

the consideration of the agent's preferences ($\overset{\pi}{\sim}_{\mathcal{T}}$). Hence, some properties introduced for CSM are revised in subjective manners and the new property is introduced.

8 CONCLUSIONS AND FUTURE WORK

This paper introduces the notion called concept similarity measure under preference profile (in symbol, $\overset{\pi}{\sim}_{\mathcal{T}}$) with the set of its desirable properties, as intended behaviors of good preference-based measures. The measure sim^{π} (cf. Section 4), which is regarded as a measure of the proposed notion, is capable of informing the degree under preferences of similarity of two concepts although they are not in the subsumption relation. At the heart of the measure is the calculation of the degrees under preferences of homomorphism between two description trees in both directions. Proofs of inherited properties are shown in Theorems 4, 5, 6, and 7. The measure can also be used regardless of the agent's preferences. Theorem 3 suggests that this is handled by the default preference profile setting, i.e. sim^{π_0}.

Apart from the mathematical definition, we suggest two concrete algorithms, viz. the top-down approach (cf. Subsection 5.1) and the bottom-up approach (cf. Subsection 5.2), for implementations of sim^{π}. The computational complexity of both algorithms is clearly discussed and is practically evaluated against $\mathcal{O}_{\text{SNOMED}}$ (cf. Subsection 6.1). The usability of possible use cases are discussed in Subsection 6.2.

The proposed measure has great potential use in knowledge engineering, such as the development of recommendation systems based on the agent's preferences, the development of domain-specific knowledge bases, and the ontology engineering. Moreover, it may be used with heterogeneous ontologies by identifying duplicated primitive concepts and primitive roles among ontologies via $\mathfrak{s}^{\mathfrak{c}}$ and $\mathfrak{s}^{\mathfrak{r}}$, respectively.

There are several possible directions for the future research. Firstly, it appears to be a natural step to extend the notion of preference profile to support more expressive DLs, e.g. universal restriction, concept negation, and also, to support an ABox. Secondly, we also aim at devising a concept similarity measure under preference profile which can handle more expressive DLs. Thirdly, we intend to explore the possibility to extend the notion of preference profile beyond $\overset{\pi}{\sim}_{\mathcal{T}}$, e.g. non-standard instance checking under preference profile. Apart from theoretical perspectives, we also intend to explore possibility on optimizing the proposed algorithmic procedures.

Acknowledgment

This research is part of the JAIST-NECTEC-SIIT dual doctoral degree program; it is supported by the Japan Society for the Promotion of Science (JSPS kaken No. 17H02258) and is partly supported by CILS of Thammasat University and the NRU project of Thailand Office of Higher Education Commission.

REFERENCES

[1] Baader, F.—Calvanese, D.—McGuinness, D. L.—Nardi, D.—Patel-Schneider, P. F.: The Description Logic Handbook: Theory, Implementation and Applications. 2nd edition. Cambridge University Press, New York, NY, USA, 2010.

[2] Ashburner, M.—Ball, C. A.—Blake, J. A.—Botstein, D.—Butler, H.—Cherry, J. M.—Davis, A. P.—Dolinski, K.—Dwight, S. S.—Eppig, J. T.—Harris, M. A.—Hill, D. P.—Issel-Tarver, L.—Kasarskis, A.—Lewis, S.—Matese, J. C.—Richardson, J. E.—Ringwald, M.—Rubin, G. M.—Sherlock, G.: Gene Ontology: Tool for the Unification of Biology. Nature Genetics, Vol. 25, 2000, No. 1, pp. 25–29, doi: 10.1038/75556.

[3] Euzenat, J.—Valtchev, P.: Similarity-Based Ontology Alignment in OWL-Lite. In: de Mántaras, R. L., Saitta, L. (Eds.): Proceedings of the 16th European Conference on Artificial Intelligence (ECAI-04), IOS Press, 2004, pp. 333–337.

[4] Janowicz, K.—Wilkes, M.: SIM-DL$_A$: A Novel Semantic Similarity Measure for Description Logics Reducing Inter-Concept to Inter-Instance Similarity. In: Aroyo, L. et al. (Eds.): The Semantic Web: Research and Applications (ESWC 2009). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 5554, 2009, pp. 353–367.

[5] Racharak, T.—Suntisrivaraporn, B.: Similarity Measures for $\mathcal{FL}_0$ Concept Descriptions from an Automata-Theoretic Point of View. Proceeding of the 2015 6th International Conference on Information and Communication Technology for Embedded Systems (IC-ICTES), 2015, pp. 1–6.

[6] D'Amato, C.—Fanizzi, N.—Esposito, F.: A Dissimilarity Measure for $\mathcal{ALC}$ Concept Descriptions. Proceedings of the 2006 ACM Symposium on Applied Computing (SAC '06), 2006, pp. 1695–1699, doi: 10.1145/1141277.1141677.

[7] Fanizzi, N.—D'Amato, C.: A Similarity Measure for the $\mathcal{ALN}$ Description Logic. Proceedings of Italian Conference on Computational Logic (CILC 2006), 2006, pp. 26–27.

[8] D'Amato, C.—Fanizzi, N.—Esposito, F.: A Semantic Similarity Measure for Expressive Description Logics. CoRR, abs/0911.5043, 2009.

[9] D'Amato, C.—Staab, S.—Fanizzi, N.: On the Influence of Description Logics Ontologies on Conceptual Similarity. In: Gangemi, A., Euzenat, J. (Eds.): Knowledge Engineering: Practice and Patterns (EKAW 2008). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 5268, 2008, pp. 48–63.

[10] Tongphu, S.—Suntisrivaraporn, B.: Algorithms for Measuring Similarity Between $\mathcal{ELH}$ Concept Descriptions: A Case Study on Snomed ct. Computing and Informatics, Vol. 36, 2017, No. 4, pp. 733–764.

[11] Baader, F.: Terminological Cycles in a Description Logic with Existential Restrictions. Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI '03), San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., 2003, pp. 325–330.

[12] Baader, F.—Brandt, S.—Küsters, R.: Matching under Side Conditions in Description Logics. Proceedings of the 17th International Joint Conference on Artificial

Intelligence – Vol. 1 (IJCAI'01), San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., 2001, pp. 213–218.

[13] Suntisrivaraporn, B.: A Similarity Measure for the Description Logic $\mathcal{EL}$ with Unfoldable Terminologies. 2013 5th International Conference on Intelligent Networking and Collaborative Systems (INCoS), 2013, pp. 408–413.

[14] Racharak, T.—Suntisrivaraporn, B.—Tojo, S.: Identifying an Agent's Preferences Toward Similarity Measures in Description Logics. In: Qi, G., Kozaki, K., Pan, J., Yu, S. (Eds.): Semantic Technology (JIST 2015). Springer International Publishing, Cham, Lecture Notes in Computer Science, Vol. 9544, 2016, pp. 201–208.

[15] Racharak, T.—Suntisrivaraporn, B.—Tojo, S.: sim^{π}: A Concept Similarity Measure under an Agent's Preferences in Description Logic $\mathcal{ELH}$. Proceedings of the 8th International Conference on Agents and Artificial Intelligence (ICAART 2016) – Vol. 2, 2016, pp. 480–487.

[16] Lehmann, K.—Turhan, A.-Y.: A Framework for Semantic-Based Similarity Measures for $\mathcal{ELH}$-Concepts. In: del Cerro, L. F., Herzig, A., Mengin, J. (Eds.): Logics in Artificial Intelligence (JELIA 2012). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 7519, 2012, pp. 307–319.

[17] Borgida, A.—Walsh, T. J.—Hirsh, H.: Towards Measuring Similarity in Description Logics. Working Notes of the International Description Logics Workshop, CEUR Workshop Proceedings, Vol. 147, 2005.

[18] Janowicz, K.: Sim-DL: Towards a Semantic Similarity Measurement Theory for the Description Logic $\mathcal{ALCNR}$ in Geographic Information Retrieval. In: Meersman, R., Tari, Z., Herrero, P. et al. (Eds.): OTM Workshops 2006. Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 4278, 2006, pp. 1681–1692, doi: 10.1007/11915072_74.

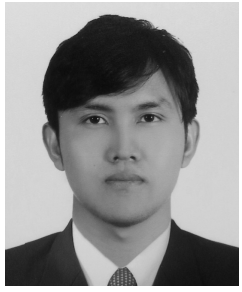
[19] Tversky, A.: Features of Similarity. Psychological Review, Vol. 84, 1977, No. 4, pp. 327–352, doi: 10.1037/0033-295X.84.4.327.

[20] Jaccard, P.: Étude Comparative de la Distribution Florale Dans Une Portion des Alpeset des Jura. Bulletin de la Societe Vaudoise des Sciences Naturellese, Vol. 37, 1901, pp. 547–579 (in French).

[21] Schulz, S.—Suntisrivaraporn, B.—Baader, F.: Snomed ct's Problem List: Ontologists' and Logicians' Therapy Suggestions. Studies in Health Technology and Informatics, Vol. 129, 2007, No. 1, pp. 802–806.

[22] Ge, J.—Qiu, Y.: Concept Similarity Matching Based on Semantic Distance. Fourth International Conference on Semantics, Knowledge and Grid, 2008, pp. 380–383, doi: 10.1109/SKG.2008.24.

[23] Giunchiglia, F.—Yatskevich, M.—Shvaiko, P.: Semantic Matching: Algorithms and Implementation. Journal on Data Semantics IX. Springer-Verlag, Berlin, Heidelberg, 2007, pp. 1–38, doi: 10.1007/978-3-540-74987-5_1.



Teeradaj Racharak received his Bachelor of Engineering with first class honors in software and knowledge engineering from Kasetsart University, Thailand, in 2010 and the Master of Engineering in computer science (with specialization in software engineering) from Asian Institute of Technology, Thailand, in 2012. Currently, he is Ph.D. student in the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), Japan, and in the School of Information, Computer, and Communication Technology (ICT), Sirindhorn International Institute of Technology (SIIT), Thammasat University, Thailand, under the JAIST-NECTEC-SIIT dual doctoral degree program. His research interest involves artificial intelligence, particularly in knowledge representation and reasoning using diverse formalisms (e.g. description logics and argumentation framework) ranging from theoretical aspects to empirical development.

Boontawee Suntisrivaraporn received his B.Eng. with first class honors in computer engineering from King Mongkut's Institute of Technology Ladkrabang, Thailand, and then graduated with both degrees of M.Sc. and D.Eng. (summa cum laude) from TU Dresden, Germany, in the field of artificial intelligence. He taught full-time at Sirindhorn International Institute of Technology and held a visiting Associate Professor position at School of Information Science, Japan Advanced Institute of Technology. His research interests mainly include description logic, knowledge representation and reasoning, biomedical ontologies, and graph theory and applications. He served as a PC member of various conferences and also received best paper awards in relevant conferences, e.g. Medinfo, ASWC, JIST. From 2016, he moved to the private sectors with his current role as Lead Data Scientist at Business Intelligence and Data Science of Siam Commercial Bank, Thailand.

Satoshi Tojo received his Bachelor of Engineering, Master of Engineering, and Doctor of Engineering degrees from the University of Tokyo, Japan. He joined Mitsubishi Research Institute, Inc. (MRI) in 1983, and the Japan Advanced Institute of Science and Technology (JAIST), Ishikawa, Japan, as Associate Professor in 1995 and became Professor in 2000. His research interest is centered on grammar theory and formal semantics of natural language, as well as logic in artificial intelligence, including knowledge and belief of rational agents. Also, he has studied the iterated learning model of grammar acquisition, and linguistic models of western tonal music.