**Abstract**

In this paper, we propose an algebraic similarity measure \( \sigma_{BS} \) (BS stands for BitSim) for assigning semantic similarity score to concept definitions in \( \mathcal{ALCH}^+ \) - an expressive fragment of Description Logics (DL). We define an algebraic interpretation function, \( I_B \), that maps a concept-definition to a unique string \((\omega_B)\) (called bit-code) over an alphabet \( \Sigma_B \) of 11 symbols belonging to \( L_B \) - the language over \( \Sigma_B \). \( I_B \) has semantic correspondence with conventional model-theoretic interpretation of DL. We then define \( \sigma_{BS} \) on \( L_B \). A detailed analysis of \( I_B \) and \( \sigma_{BS} \) has been given.

1 Introduction

Semantic similarity measure serves as the foundation of knowledge discovery and management processes such as ontology matching, ontology alignment & mapping, ontology merging, etc [Shvaiko and Euzenat, 2013]. Ontological concept similarity can be based on different approaches: (i) string matching of concept labels (i.e. lexical similarity) [Stoilos et al., 2005], (ii) external lexical resource/ontology based matching (i.e. lexicosemantic similarity) [Rada et al., 1989a], (iii) graph-based matching using lexicons such as WordNet [Stuckenschmidt, 2007] (i.e. structural similarity), (iv) property analysis (as in FCA-based similarity [Cimiano et al., 2005]) or instance analysis (as in Jaccard similarity [Jaccard, 1998]) based matching over a large sample of concept instance occurrences (i.e. instance-driven similarity), (v) matching based on statistical analysis of attribute-value or distribution analysis within fixed context-windows of concepts over large corpora (i.e. statistical similarity) [Li and Clifton, 1994], and (vi) model-theoretic matching of formal concept descriptions (i.e. formal semantic similarity) [Alsubait et al., 2014].

It can be argued that, in comparison to other approaches, formal semantic similarity measure modeling has not received equal research attention. Nevertheless, existing literature is significant, and can be broadly classified into two approaches: (i) Propositional Logics based [Nienhuys-Cheng, 1998; Ramon and Bruynooghe, 1998], and (ii) Description Logics (DL) based [Alsubait et al., 2014; Lehmann and Turhan, 2012; Stuckenschmidt, 2007; Fanizzi and dAmato, 2006; Borgida et al., 2005]. The former requires: (a) representation of ontologies (mostly in RDFS/OWL format) in First Order Predicate Logic, (b) a set of axioms (or domain knowledge, mostly as upper ontologies/thesaurus), and (c) a SAT solver that checks satisfiability (and hence, satisfiability) of disjointness of concept pairs. The latter approach, on the other hand, does not necessarily require any formal language transformation or satisfiability checker. In this paper we propose an algebraic similarity measure, called BitSim (\( \sigma_{BS} \)), that can compute semantic similarity of pair of concepts defined in \( \mathcal{ALCH}^+ \). The motivation behind \( \sigma_{BS} \) is to formulate a formal semantic similarity measure that provides: (i) a platform for fast, scalable, and accurate semantic similarity computation of DL concepts, and (ii) a sound and complete correspondence with conventional semantic interpretation of DL. \( \sigma_{BS} \) is algebraic, in the sense that it maps a given pair of concept codes (called bit-code), instead of concept DL definitions/axioms, to a positive real space. For this we define a novel algebraic interpretation function, called \( I_B \), that maps an \( \mathcal{ALCH}^+ \) definition to a unique string, called bit-code, \((\omega_B)\) belonging to the language \( L_B \) defined over a novel algebraic alphabet \( \Sigma_B \). We prove that \( I_B \) has complete correspondence with \( I_{\mathcal{ALCH}^+} \). We also show that \( \sigma_{BS} \) is highly adaptive to any kind of similarity measure that relies on set operation. As an example, we have shown how \( \sigma_{BS} \) can be plugged into Jaccard similarity index. The contribution of the paper is as follows:

- \( I_B \) : A novel algebraic semantic interpretation function for \( \mathcal{ALCH}^+ \).
- Proof of mathematical correspondence of \( I_B \) with semantic interpretation of \( \mathcal{ALCH}^+ \).
- \( \sigma_{BS} \) : A novel semantic similarity measure based on \( I_B \).
- Comparative analysis of properties of \( \sigma_{BS} \) with contemporary DL based similarity measures.

2 Related Work

DL based similarity measures, as described in the introduction, can be further sub-divided into: (i) taxonomic analysis, (ii) structural analysis [Tongphu and Sunitsrivarporn, 2014; Ontaño and Plaza, 2012] [Joslyn et al., 2008; Hariri et al., 2006], (iii) language approximation [Stuckenschmidt, 2007].
Tserendorj et al., 2008, Groot et al., 2005, Brandt et al., 2002, and (iii) model-theoretic analysis [Distel et al., 2014; Alsubait et al., 2014; Lehmann and Turhan, 2012; Borgida et al., 2005]. The most common approach for DL based similarity measure modeling adopts taxonomic analysis as proposed in [Rada et al., 1989b; Resnik and others, 1999; Jiang and Conrath, 1997] [Wu and Palmer, 1994; Lin, 1998]. These techniques can be further sub-divided, as mentioned in introduction, into graph-traversal approaches [Rada et al., 1989b] and Information-Content approaches [Resnik and others, 1999]. However, these methods can work on a generalized ontology and hence, are not sensitive to DL definitions.

In structural analysis based approaches, a similarity measure is designed to capture the semantic equivalence of description trees of definitions of DL concept pairs. One way of achieving this is to calculate the degree of homomorphism between such trees, as proposed in [Tongphu and Suntisrivaraporn, 2014]. A refinement graph based anti-unification approach has been proposed in [Ontaño and Plaza, 2012] for computing instance similarity. Approximation based techniques aim at converting given DL expression to another lower expressive DL language on approximation. In [Stuckenschmidt, 2007] an upper and lower approximation interpretation for SHIQ have been defined over a sub-vocabulary of SHIQ. The sub-vocabulary can be formed by either removing simpler concepts in a given definition or by replacing them with structurally simpler concepts [Groot et al., 2005]. Another technique, as proposed in [Noia et al., 2004], is based on converting user query into a DL expression and try to classify the match to be either an exact match, or a potential match (i.e., match might happen if some concept atom and operators are added) or a partial match (where the user query and answer/description found in the knowledge base are in conflict).

One of the pioneer work on model-theoretic interpretation based approach can be found in [Borgida et al., 2005]. The work shows the inherent difficulty in measuring similarity of DL concepts using conventional taxonomic analysis based techniques. It then uses an Information-Content based approach to evaluate the similarity of two concept definitions. A work has been proposed by [Lehmann and Turhan, 2012] for similarity computation of concepts defined in ELH. In this work, a Jaccard Index [Jaccard, 1998] based approach has been followed that compares common parents of a concept pair using a fuzzy connector (i.e. a similarity score aggregation function). A similar Jaccard Index based approach has been recently proposed in [Alsubait et al., 2014]. Another recent approach has been proposed in [Distel et al., 2014]. The work emphasizes the necessity of triangle inequality property of formal semantic similarity measure. It defines two versions of a relaxation function for computing dissimilarity of concepts defined in EL. However, it can be proved that triangle inequality is not always valid and hence, is not a necessary condition.

3 Preliminaries

3.1 \( \mathcal{ALCH}^+ \) - A Description Logics Fragment

We hereby define the semantic interpretation of \( \mathcal{ALCH}^+ \) (an extension of which also interprets current OWL 2 specification). Let \( \mathcal{I} \) be an interpretation function, and \( \Delta \) be the universal domain. \( \mathcal{ALCH}^+ \) is defined as:

- **\( \mathcal{AL} \) (Attributive Language):**
  - Atomic concept: \( A^I \subseteq \Delta^I \)
  - Role: \( r^I \subseteq \Delta^I \times \Delta^I \)
  - Atomic Negation: \( \neg A^I = \Delta^I \setminus A^I \)
  - Top Concept: \( \top = \Delta^I \)
  - Bottom Concept: \( \bot = \emptyset^I \)
  - Conjunction: \( (C \cap D)^I = C^I \cap D^I \)
  - Value Restriction: \( (\forall R.C)^I = \{ a \mid \forall b.(a,b) \in R^I \Rightarrow b \in C^I \} \)
  - Limited Existential Restriction: \( (\exists R.C)^I = \{ a \mid \exists b.(a,b) \in R^I \land b \in C^I \} \)

- **\( \mathcal{E} \) (Full Existential Restriction):** \( (\exists R.C)^I = \{ a \mid \exists b.(a,b) \in R^I \land b \in C^I \} \)

- **\( \mathcal{C} \) (Concept Negation):** \( \neg C^I = \Delta^I \setminus C^I \)

- **\( \mathcal{H} \) (Role Hierarchy):** \( R_1^I \subseteq R_2^I \)

- **\( \mathcal{R}_{\text{UNION}} \) (Role Union):** \( R_1^I \cup R_2^I \)

- **\( \mathcal{R}_{\text{INTERSECTION}} \) (Role Intersection):** \( R_1^I \cap R_2^I \)

3.2 Formal Similarity Measure

In this section we define the algebraic properties of \( \sigma \) as given in [Lehmann and Turhan, 2012]. Let \( C \) is a DL concept in a given terminology (T-Box) \( T \).

**Definition 1:** A semantic similarity measure \( \sigma \) is a function defined as follows:

\[ \sigma : C \times C \mapsto [0, 1] \text{ where } C \in T \]

**Properties of Similarity Measure:** Arguably \( \sigma \) should hold the following properties:

- **Positiveness**: \( \forall C_i, C_j \in T, \sigma(C_i, C_j) \geq 0 \)  \( \text{(1)} \)
- **Reflexive**: \( \forall C_i \in T, \sigma(C_i, C_i) = 1 \)  \( \text{(2)} \)
- **Maximality**: \( \forall C_i, C_j, C_k \in T, \sigma(C_i, C_k) \geq \sigma(C_i, C_j) \)  \( \text{(3)} \)
- **Symmetry**: \( \forall C_i, C_j \in T, \sigma(C_i, C_j) = \sigma(C_j, C_i) \)  \( \text{(4)} \)
- **Equivalence Closure**: \( \forall C_i, C_j \in T, \sigma(C_i, C_j) = 1 \)
  \[ \iff C_i \equiv C_j \]  \( \text{(5)} \)
- **Equivalence Invariance**: \( \forall C_i, C_j, C_k \in T, C_i \equiv C_j \)
  \[ \Rightarrow \sigma(C_i, C_k) = \sigma(C_j, C_k) \]  \( \text{(6)} \)

\( \text{\textsuperscript{2}}\) The syntax of \( \mathcal{ALCH}^+ \) follows conventional DL as defined in [Baader, 2003].

\( \text{\textsuperscript{*}}\) denotes that the property is not universally adopted as necessary condition. Also it cannot be proven to be valid in all types of algebraic spaces.
Structural Dependency: \( \forall C_n, n, C, C_i \in T, \exists C_{\leq n} = \bigcap_{k<\infty} (C_k \cap C_i) \); and, \( C_{\leq n} = \bigcap_{k<\infty} (C_k \cap C_i) \);

then, \( \lim_{n \rightarrow \infty} \sigma(C_{\leq n}) = 1 \) (7)

Subsumption Preservation: \( \forall C_i, C_j, C_k \in T, \)

\( C_i \subseteq C_j \subseteq C_k \implies \sigma(C_i, C_j) \geq \sigma(C_i, C_k) \) (8)

Reverse Subsumption Preservation: \( \forall C_i, C_j, C_k \in T, \)

\( C_i \subseteq C_j \subseteq C_k \implies \sigma(C_i, C_k) \geq \sigma(C_i, C_j) \) (9)

Strict Monotonicity: \( \forall C_i, C_j, C_k \in T, \exists C_i \subseteq C_j \subseteq C_k \implies \sigma(C_i, C_j) \geq \sigma(C_i, C_k) \)

And, \( \forall C_i \subseteq C_j \subseteq C_k \) and \( C_j \subseteq C_m \)

then, \( \sigma(C_i, C_k) \geq \sigma(C_i, C_j) \) (10)

It should be noted that the aforementioned necessary properties may not be sufficient and hence, detailed theoretical analysis has to be done on sufficiency.

4 \( L_B \): Formal Language for Concept Coding

In this section we define the formal language \( L_B \) on which the proposed algebraic interpretation function \( f_B \) is defined. We first define bit (\( \Sigma_B \)), the alphabet of \( L_B \), as follows:

**Definition 2:** A base alphabet (\( \Sigma_{\text{base}} \)) is an alphabet defined as: \( \Sigma_{\text{base}} = \{0, 1\} \)

- 0 is the empty symbol. It is also called potential bit since it generates all other symbols (i.e. bits).
- 1 is base bit, called property bit, signifying the presence of a property at the string position that it holds.

**Definition 3:** An bit operator (\( \oplus \)) is a set of operators on the base bits defined as: \( \oplus_B = \{ \oplus_B, \ominus_B, \neg_B \} \).

We now define a very important semantics for potential bit (i.e. 0) as follows:

\( \neg_B 0 = 0 \) (1)

**Definition 4:** An derived alphabet (\( \Sigma_{\text{derived}} \)) is an alphabet defined as: \( \Sigma_{\text{derived}} = \{ X, X'', X', Y', \top', \bot', \top, \bot \} \)

where

- \( X'' = \neg_B 1 \)
- \( X = 1 \ominus_B 0 \)
- \( X' = X'' \ominus_B 0 \)
- \( \top' = Y \ominus_B Y' \)
- \( Y' = 1 \ominus_B 0 \)
- \( Y = X'' \ominus_B 0 \)
- \( \bot' = X \ominus_B X' \)

Based on the above definition the following observations can be made (using de Morgan’s law):

- \( X = \neg_B Y \)
- \( X' = \neg_B Y' \)
- \( \top' = \neg_B \bot' \)
- \( \bot = \neg_B \top' \)

A further analysis shows that \( \Sigma_B \) has a partial ordering \( \preceq_B \) (as shown in figure 1). It is interesting to note that \( b_i \ominus_B \neg_B b_i \neq \bot, \forall b_i \in \Sigma_B \).

**Definition 5.1:** An bit-alphabet (\( \Sigma_B \)) is defined as \( \Sigma_B = \Sigma_{\text{base}} \cup \Sigma_{\text{derived}} \).

It is to be noted that \( \oplus_B \) satisfies commutativity and double negation over \( \Sigma_B \).

**Definition 5.2:** A quantifier-alphabet (\( \Sigma_{\text{q-}} \)) is defined as \( \Sigma_{\text{q-}} = \{ 1, 0, X, Y \} \). The algebraic space of \( \Sigma_{\text{q-}} \) is defined as below:

- \( X = 1 \ominus_B 0 \)
- \( Y = 1 \ominus_B 0 \)
- \( \neg_B 1 = 0 \)
- \( \neg_B Y' = X \)

**Definition 5.3:** A role-alphabet (\( \Sigma_{\text{r-}} \)) is defined as \( \Sigma_{\text{r-}} = \{ 1, 0, X, Y \} \). The algebraic space of \( \Sigma_{\text{r-}} \) is defined as below:

- \( X = 1 \ominus_B 0 \)
- \( Y = 1 \ominus_B 0 \)

**Definition 6.1:** A base bit-code (\( \omega_{\text{base}} \)) is defined as \( \omega_{\text{base}} = 0^* \oplus_B b_1^* \ominus_B (\ominus_B b_2 b_3) \).

**Definition 6.2:** A derived bit-code (\( \omega_{\text{derived}} \)) is defined as \( \omega_{\text{derived}} = (\omega_{\text{derived}}) \ominus_B (\omega_{\text{derived}}) \).

It can be observed that the definition of \( \omega_{\text{derived}} \) is recursive. We leave the explanation and utility of the definition in section 5.3.

**Definition 7:** \( L_B \) is defined as \( L_B = \{ \omega_{\text{base}} \cup \omega_{\text{derived}} \} \).

5 Encoding \( \mathcal{ALC} + \) Concept

5.1 Motivation

The motivation behind BitSim (\( \sigma_{\text{BS}} \)) is to develop a formal, efficient, and scalable matchmaking system that can be applied in DL based knowledge bases. Unlike other DL based similarity measures, \( \sigma_{\text{BS}} \) was designed to satisfy all the necessary properties defined in section 3.2 with special emphasis on structural dependency and strict monotonicity.
At the same time, $\sigma_{BS}$ computation is over $L_B$, rather than $\mathcal{ALCH}^+$. This significantly improves the computational speed since $\sigma_{BS}$ essentially becomes a function over $\omega_g$ pairs (such bit-codes can be computed and stored off-line in the knowledge base). Since, $\sigma_{BS}$ computation is performed on pairs of bits holding the same position in $\omega_g$, therefore we can chunk bit-codes in constant sizes and perform similarity over concept pairs on parallel computational platforms. This gives massive scalability to $\sigma_{BS}$. Efficient optimization can be performed by caching similarity results of bit-code chunks that are frequently visited.

We will also show that, at a bit level, $\sigma_{BS}$ has a partial ordering $\preceq_{BS}$. This allows application-oriented assignment of similarity scores to bit pairs at the lowest granularity. Also, $\sigma_{BS}$ is highly adaptive to all types of similarity measures that have set theoretic operations on DL concepts.

5.2 Encoding Atomic Concepts

Before we show that $L_B$ has complete correspondence with $\mathcal{ALCH}^+$, we first provide the foundational axioms that helps us to encode atomic concepts in $\mathcal{ALCH}^+$. For that we need to define the proposed algebraic interpretation function $I_B$ (also called bit-interpretation).

**Definition 8:** Bit-interpretation ($I_B$) is a function as follows:

$$I_B : C \rightarrow L_B : C \in \mathcal{T}_{\mathcal{ALCH}^+}$$

We hereby define $(A_i)^s$, where $A_i$ is an arbitrary atomic concept in $\mathcal{ALCH}^+$, using the following two axioms:

**Fundamental Axiom of Atomic Concepts**

$$\forall b^A_k \in \omega^A_B, b^A_k \in \Sigma_{BS}, \omega^A_B \equiv (A_i)^s$$

**Axiom of Significant Property Bit**

$$\exists b^A_k \in \omega^A_B, b^A_k = 1 \& b^A_{k+1} = 0^s$$

where, $k$ is the $k$-th position (in increasing order from right to left) in $\omega^A_B$. In the second axiom $b^A_k$ holds the significant property bit.

We now show the method to encode inclusion axioms on atomic concepts using the following two axioms:

**Axiom of Property Bit Inheritance**

$$\forall A_i, A_j \in \mathcal{T}_{\mathcal{ALCH}^+}, A_i \subseteq A_j, b^{A_j}_k = 1 \rightarrow b^{A_i}_k = 1; \forall b^A_k \in \omega^A_B$$

5.3 Encoding Derived Concepts

For encoding derived concepts, we cannot attain completeness using $\omega_{BS}$. This is because, for a bounded number of atomic concepts (say, $n$) we need a mechanism to encode $2^n$ distinct and disjoint concepts, in the worst case. However, with only 11 bits in $\omega_g$, we can only encode 11 distinct concepts. It is because of this reason that we need to use $\omega_{BS}$. The method is to have nested encoding for bit operations over certain special bit pairs. These operations do not have direct mapping to the algebraic lattice shown in figure 1. Instead they map to intermediate and discrete compound bits ($b_ω$), which can be represented in terms of $\Sigma_ω$. We define $b_ω$ as follows:

**Definition 9:** A compound bit is defined as:

$$b_ω = (ω_{BS} \omega_{BS})$$

where the algebra of $b_ω$ is defined as shown in figure 3.

For any derived concept $C_i$, we can state the following:

**Axiom of Binary Concept Operation**

$$b^{C_i} \circ b^{C_j} \equiv ((\omega_{BS} \omega_{BS}) \circ b^{C_i}) \circ (\omega_{BS} \omega_{BS}) b^{C_j} \equiv b^{C_i \circ C_j}$$

Based on axiom 5 we can state that:

**Lemma 1.** For any binary operation between $b^{C_k}$ and $b^{C_l}$, the length of both the operands, $b^{C_k}$ and $b^{C_l}$, must be same; where $k$: $k$-th position of $b^{C_k}$ in $ω_g$.

**Lemma 2.** For any binary operation between $b^{C_k}$ and $b^{C_l}$, the length of the resultant $b^{C_k \circ C_l}$ has growth $= O(c^n)$, where $c \in \{1, 2, 3\}$ and $n$ is length of operand $b_ω$.

**Lemma 3.** If $\exists b_k \in \omega_g$ holds $b_k = T$, then $\omega_g = T$

**Lemma 4.** If $\exists b_k \in \omega_g$ holds $b_k = \bot$, then $\omega_g = \bot$

**Lemma 5.** If $\forall b_k \in \omega_g$ holds $b_k = T'$, then $\omega_g = T'$

**Lemma 6.** If $\forall b_k \in \omega_g$ holds $b_k = \bot'$, then $\omega_g = \bot'$
We now postulate the following axioms for operations over derived concepts:

**Axiom of Concept Negation**

\[(\neg C)_f^S \equiv \{ \neg b^C_k \} \equiv \{ b_k^C \} \] (6)

**Axiom of Binary Concept Operation**

\[(C_i \oplus C_j)_f^S \equiv (C_i)_f^S \oplus (C_j)_f^S \equiv \{ b_k^C_i \} \oplus \{ b_k^C_j \} \] (7)

\[\equiv \{ b_k^{C_i \oplus C_j} \}; \text{where } \oplus \in \{ \lor, \land \} \]

**Axiom of Universal Role Restriction**

\[(\forall R_i C_j)_f^S \equiv ((\forall \omega b^R_i) \circ (\forall \omega b^R_j) \circ) \circ \omega_S \{ b_k^C_i \} \] (8)

**Axiom of Full Existential Role Restriction**

\[(\exists R_i C_j)_f^S \equiv ((\exists \omega b^R_i) \circ (\exists \omega b^R_j) \circ) \circ \omega_S \{ b_k^C_i \} \] (9)

It is to be noted that for the above axioms \( b_k \) can be both a simple and a compound bit. We now provide a proof for showing the mathematical correspondence between \((L_B f^S) \) and \((\mathcal{ALCH}^+ f^S) \).

**Lemma 7.** \( \forall b^C \in \omega^C, b^C_k \in \omega^C; \ k: k\text{-th position in } \omega; \ b^C_k \preceq b^C_k \iff (C)_f^{\mathcal{ALCH}}, \subseteq (C)_f^{\mathcal{ALCH}} \).

**Proof.** Proof can be derived from the lattice structure of \( \Sigma_B \) (see figure 1.)  \(\square\)

Following the above lemma we can state that:

**Theorem 1.** \( (C_i)_f^S \subseteq (C_i)_f^S \iff (C_i)_f^S \subseteq (C_i)_f^S \)

**Theorem 2.** \( (C_i \oplus C_j)_f^{\mathcal{ALCH}} \iff (C_i \oplus C_j)_f^{\mathcal{ALCH}} \)

**Proof.** \( \forall b^C_k \in \omega^C, b^C_k \in \omega^C; \ b^C_k \preceq b^C_k \iff b^C_k \preceq b^C_k \iff b^C_k \preceq b^C_k \), if \( \preceq \in \mathcal{ALCH} \); else converse \(\{ \preceq \in \mathcal{ALCH} \).  \(\square\)

**Theorem 3.** \( (QR.C)_f^S \iff (QR.C)_f^{\mathcal{ALCH}}; \ Q \in \{ \forall, \exists \} \)

**Proof.** From theorems 7 and 8, we can show that \((QR.C)_f^{\mathcal{ALCH}} \) is unique. This is because \( Q \) is unique. In other words, the algebra has complete correspondence with \((QR.C)_f^S \).  \(\square\)

**Theorem 4.** \( \forall \omega_S \in \mathcal{L}_B ; \omega_S \) is unique.

**Proof.** Follows from axiom 4 and lemma 1 - 6.  \(\square\)

**Theorem 5.** \( \mathcal{ALCH}^+ f^S \) has complete correspondence with \( L_B f^S \)

**Proof.** The proof follows from theorem 1 - 3.  \(\square\)

### 6 BitSim Similarity Measure

#### 6.1 Outline

In this section we provide a generic definition for BitSim. We first define \( \sigma_{BS} \) (i.e. similarity at a bit level) as follows:

**Definition 10:** \( \sigma_{BS} : b_i \times b_j \mapsto [0, 1] \); where \( b_i, b_j \in \Sigma_B \).

Is to be noted that \( b_i \) can be \( b_{as} \) as well. One can see that \( \sigma_{BS} \) has a total order \( \preceq_{BS} \) (see figure 4). In order to compute similarity at a bit-code level, we define an aggregation function called \( \hat{\sigma}_{BS} \). There are two parameters that should influence the value output of \( \hat{\sigma}_{BS} \): (i) \( \sigma_{BS} \) and (ii) code-generativity (CG). We define code-generativity as follows:

**Definition 11:** Code-generativity \((\mathcal{CG}) \) of any \( \omega_B \) of a concept is the total number of distinct and disjoint concepts that are covered by \( \omega_B \).

As an example, \((\mathcal{CG})(\{X1\}) = 3 \); (i.e. 11, 111, 111). We now define the similarity measure at a bit-code level (we call it \( \sigma_{BS} \)) as follows:

**Definition 12:** \( \hat{\sigma}_{BS} : [\sigma_{BS} \times (\mathcal{CG}) \times \preceq_{BS}] \mapsto [0, 1] \).

We now postulate the following axioms:

**Axiom of 0-bit Similarity**

\[ \forall k; \sigma_{BS}(0_k, 0_k) \text{ is ignored.} \] (10)

**Axiom of } \neg \text{ bit Similarity**

\[ \forall a \in \Sigma_B; \sigma_{BS}(\top, a) = 1 \text{ if } a = \top; \text{ else undefined.} \] (11)
6.2 \( \sigma_{BS} \): Property Analysis

In this section we show that \( \sigma_{BS} \) follows the necessary conditions: (i) reflexive, (ii) maximality, (iii) equivalence closure, (iv) equivalence invariance, (v) structural dependency, (vi) subsumption preservation, (vii) reverse subsumption preservation, and (viii) strict monotonicity. The first two properties trivially hold true. The following theorems show that the rest of the properties hold:

**Theorem 6.** Equivalence closure and invariance holds true for \( \sigma_{BS} \).

**Proof.** Follows from theorem 1 and theorem 4. \( \square \)

**Theorem 7.** Structural dependency holds true for \( \sigma_{BS} \).

**Proof.** Under the condition of structural dependency, for two concepts \( C_i \) and \( C_j \) the \( \omega_2 \) that is generated for them will have a length, say \( l \), that grows exponentially as the number of inner intersections in the definition of both \( C_i \) and \( C_j \) tends to \( \infty \). Therefore, except for the word length of the invariant concepts in the definition of \( C_i \) and \( C_j \), all of \( \omega_2 1 \) and \( \omega_2 C_i \) will be exact. Hence, \( \sigma_{BS}(C_i, C_j) \rightarrow 1 \) \( \square \)

**Theorem 8.** Strict monotonicity holds true for \( \sigma_{BS} \).

**Proof.** When more than one concepts (say, \( C_x, C_y \)) subsume two (say, \( C_i, C_j \)) out of three arbitrary concepts (say, \( C_i, C_j, C_k \)), while only one (say, \( C_x \)) subsumes all three, then \( \sigma_{BS}(C_i, C_j) > \sigma_{BS}(C_i, C_k) \). This is because, since \( \sigma_{BS} \) is property based measure, \( C_i, C_j \) will inherit more common bits than \( C_k \). \( \square \)

We now show how \( \sigma_{BS} \) can be adapted to a third-party similarity measure as follows:

**Definition 13:** \( \sigma_{BS}-Jaccard = \sigma_{BS}(\omega_{BS}^{C_i/C_j}, \omega_{BS}^{C_i/C_k}) \)

7 Discussion

As can be seen, \( \sigma_{BS} \) can be adapted as an alternate paradigm for DL subsumption reasoning. Since \( L_B \) can be mapped to a boolean space one can perform bit operations at high speed and that too on a distributed and parallel platform. At the same time, various caching techniques can be applied efficiently. In the future we will be exploring these research prospects and other possibilities such as probabilistic reasoning on \( L_B \), A-Box reasoning, and reasoning over higher expressive DL.

8 Conclusion

In this paper we have proposed BitSim (\( \sigma_{BS} \)) - an algebraic similarity measure for concept definitions in \( \mathcal{ALCH}^+ \). We show that \( \sigma_{BS} \) satisfies all the necessary algebraic properties recommended for a formal similarity measure. Being based on \( I_B \), \( \sigma_{BS} \) is highly sensitive to standard DL interpretation. Furthermore, \( \sigma_{BS} \) is highly adaptive to any similarity measure that uses set theoretic operations.

References

[Alsubait et al., 2014] Tahani Alsubait, Bijan Parsia, and Uli Sattler. Measuring similarity in ontologies: A new family of measures. In Knowledge Engineering and Knowledge Management, pages 13–25. Springer, 2014.

[Baader, 2003] Franz Baader. The description logic handbook: theory, implementation, and applications. Cambridge university press, 2003.

[Borgida et al., 2005] Alexander Borgida, Thomas J Walsh, and Haym Hirsh. Towards measuring similarity in description logics. Description Logics, 147, 2005.

[Brandt et al., 2002] Sebastian Brandt, Ralf Küsters, and Anni-Yasmin Turhan. Approximation and difference in description logics. In KR, pages 203–214, 2002.

[Cimiano et al., 2005] Philipp Cimiano, Andreas Hotho, and Steffen Staab. Learning concept hierarchies from text corpora using formal concept analysis. J. Artif. Intell. Res.(JAIR), 24:305–339, 2005.

[Distel et al., 2014] Felix Distel, Jamal Atif, and Isabelle Bloch. Concept dissimilarity with triangle inequality. KRI4, 2014.

[Fanizzi and dAmato, 2006] Nicola Fanizzi and Claudia dAmato. A similarity measure for the \( \mathcal{ALN} \) description logic. Proceedings of CILC, pages 26–27, 2006.

[Groot et al., 2005] Perry Groot, Heiner Stuckenschmidt, and Holger Wache. Approximating description logic classification for semantic web reasoning. In The Semantic Web: Research and Applications, pages 318–332. Springer, 2005.

[Hariri et al., 2006] Babak Bagheri Hariri, Hassan Abolhassani, and Ali Khodaei. A new structural similarity measure for ontology alignment. In 3WWS, pages 36–42, 2006.

[Jaccard, 1998] P Jaccard. Etude comparative de la distribution florale dans une portion des alpes et des jura. In Bull Soc Vaudoise Sc Nat, volume 37, pages 547–579, 1998.

[Jiang and Conrath, 1997] Jay J Jiang and David W Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008, 1997.
[Joslyn et al., 2008] Cliff Joslyn, Alex Donaldson, and Patrick R Paulson. Evaluating the structural quality of semantic hierarchy alignments. In International Semantic Web Conference (Posters & Demos), volume 401, 2008.

[Lehmann and Turhan, 2012] Karsten Lehmann and Anni-Yasmin Turhan. A framework for semantic-based similarity measures for ELC-concepts. In Logics in Artificial Intelligence, pages 307–319. Springer, 2012.

[Li and Clifton, 1994] Wen-Syan Li and Chris Clifton. Semantic integration in heterogeneous databases using neural networks. In VLDB, volume 94, pages 12–15, 1994.

[Lin, 1998] Dekang Lin. An information-theoretic definition of similarity. In ICML, volume 98, pages 296–304, 1998.

[Nienhuys-Cheng, 1998] Shan-Hwei Nienhuys-Cheng. Distances and limits on herbrand interpretations. In Inductive Logic Programming, pages 250–260. Springer, 1998.

[Noia et al., 2004] Tommaso Di Noia, Eugenio Di Sciascio, Francesco M Donini, and Marina Mongiello. A system for principled matchmaking in an electronic marketplace. International Journal of Electronic Commerce, 8(4):9–37, 2004.

[Ontañón and Plaza, 2012] Santiago Ontañón and Enric Plaza. Similarity measures over refinement graphs. Machine learning, 87(1):57–92, 2012.

[Rada et al., 1989a] Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. Development and application of a metric on semantic nets. Systems, Man and Cybernetics, IEEE Transactions on, 19(1):17–30, 1989.

[Rada et al., 1989b] Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. Development and application of a metric on semantic nets. Systems, Man and Cybernetics, IEEE Transactions on, 19(1):17–30, 1989.

[Ramon and Bruynooghe, 1998] Jan Ramon and Maurice Bruynooghe. A framework for defining distances between first-order logic objects. In Inductive Logic Programming, pages 271–280. Springer, 1998.

[Resnik and others, 1999] Philip Resnik et al. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J. Artif. Intell. Res.(JAIR), 11:95–130, 1999.

[Shvaiko and Euzenat, 2013] Pavel Shvaiko and Jérôme Euzenat. Ontology matching: state of the art and future challenges. Knowledge and Data Engineering, IEEE Transactions on, 25(1):158–176, 2013.

[Stoilos et al., 2005] Giorgos Stoilos, Giorgos Stamou, and Stefanos Kollias. A string metric for ontology alignment. In The Semantic Web–ISWC 2005, pages 624–637. Springer, 2005.

[Stuckenschmidt, 2007] Heiner Stuckenschmidt. Partial matchmaking using approximate subsumption. In Proceedings of the national conference on Artificial Intelligence, page 1459, 2007.

[Tongphu and Suntisrivaraporn, 2014] Suwan Tongphu and Boontawee Suntisrivaraporn. On desirable properties of the structural subsumption-based similarity measure. 4th Joint International Semantic Technology Conference (JIST), 2014.

[Tserendorj et al., 2008] Tuvshintur Tserendorj, Sebastian Rudolph, Markus Krötzsch, and Pascal Hitzler. Approximate owl-reasoning with screech. In Web Reasoning and Rule Systems, pages 165–180. Springer, 2008.

[Wu and Palmer, 1994] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pages 133–138. Association for Computational Linguistics, 1994.