A logic-based framework for characterizing nexus of similarity within knowledge bases (Extended Abstract)

Giuseppe Agresta[1] , Giovanni Amendola[1] , Pietro Cofone[1] , Marco Manna[1] , Aldo Ricioppo[1,2]
[1]Department of Mathematics and Computer Science, University of Calabria
[2]Department of Computer Science, University of Cyprus
{name.surname}@unical.it

Context. Similarities play a key role in many real-world scenarios, driving extensive research into methodologies for measuring entity similarity and expanding sets of entities with similar ones. Machines can nowadays deal with these tasks by taking in some regard relevant interconnected properties shared by entities, which we term *nexus of similarity*.

Researchers from various fields have proposed a range of approaches to measure the *semantic similarity* between entities (Gomaa and Fahmy 2013). For example, modern machines are capable of computing a plausibly high similarity score between $\langle\mathsf{Paris}\rangle$ and $\langle\mathsf{Rome}\rangle$, by taking into account somehow that both of them are “European cities”, “places situated on rivers”, “capitals”, and so on.

Inspired by “Google Sets” (Cirasella 2007), considerable academic and commercial efforts have been also devoted to providing solutions for expanding a given set of entities with similar ones. The main tasks are: *entity set expansion*, *entity recommendation*, *tuples expansion*, or *entity suggestion*. For example, one can expand the set $\mathbf{U} = \{\langle\mathsf{Paris}\rangle, \langle\mathsf{Rome}\rangle\}$ and obtain $\mathbf{U}' = \mathbf{U} \cup \{\langle\mathsf{Amsterdam}\rangle\}$; then, one can reapply the process starting from $\mathbf{U}'$ to obtain the set $\mathbf{U}'' = \mathbf{U}' \cup \{\langle\mathsf{Brussels}\rangle, \langle\mathsf{Rio\ de\ Janeiro}\rangle, \langle\mathsf{Vienna}\rangle\}$. Indeed, all these entities share one or more of the aforementioned properties, e.g., “European cities”, “places situated on rivers”.

Complementary approaches, ranging from DLs (Cohen, Borgida, and Hirsh 1992) to Semantic Web (Colucci et al. 2016; Petrova et al. 2019) and Database Theory (ten Cate et al. 2023), studied the task of recognizing and formally expressing/explaining nexus of similarity (a.k.a. commonalities) between entities within a Knowledge Base (KB).

Motivation. The above approaches vary across some key dimensions: the form of the input, (e.g., pairs of entities, sets of entities, sets of entity tuples); the type of KBs they can handle, (e.g., DL-KBs, RDF documents, DBs, or even text corpora); the portion of knowledge used to describe the input, (e.g., the entire KB or selected excerpts); and the specific formalism to express commonalities (e.g., DL-Concepts, r-graphs, (U)CQs, SPARQL, rooted-CQs). Also, commonalities may not be finitely expressible in some setting, and expansions are generally viewed as “linear” (e.g., $\mathbf{U} \subset \mathbf{U}' \subset \mathbf{U}''$), rather than “taxonomic” (e.g., $\mathbf{U} \subset \mathbf{U}'$ and $\mathbf{U} \subset \mathbf{U}''$, where $\mathbf{U}'$ and $\mathbf{U}''$ are not comparable under subset inclusion). Thus, a unifying framework is missing.

Contribution. Amendola, Manna, and Ricioppo (2024) proposed a general logic-based framework for characterizing (i.e., explaining/expressing in a comprehensive way) nexus of similarity, between entity tuples, within KBs. The paper introduced the notion of *selective KB*, denoted by $\mathcal{S} = (K, \varsigma)$, to enhance any KB K (possibly beyond DL/RDF) with a *summary selector* ς: basically, ς is a function that, for any tuple τ of entities, selects a relevant portion of the knowledge entailed by K that describes τ. Then, they designed a suitable *nexus explanation language*, called $\mathbb{NC}\mathrm{F}$, and equipped it with an appropriate semantics. Accordingly, the paper defines $\mathbb{NC}\mathrm{F}$-formulas —playing the role of *explanations* and *(canonical/core) characterizations*— to express nexus of similarity between tuples of entities within $\mathcal{S}$, demonstrating that these formulas always exist and are computable. In particular, core characterizations are not only comprehensive, but also concise and human understandable. Furthermore, they introduced the *expansion graph*, generalizing the classical notion of linear expansions. The work also proposed and studied key reasoning tasks related to the computation of characterizations and expansions, showing tractability under practical assumptions.[1]

Framework overview. Consider the knowledge graph $\mathcal{G}_0$ in Figure 1. It can be naturally encoded as the dataset:

$$D_0 = \{\mathsf{isa}(\mathsf{Epcot}, \mathsf{tp}), \mathsf{located}(\mathsf{Epcot}, \mathsf{Florida}), \ldots\}.$$

Given an ontology $O_0 = \{\mathsf{isa}(x,z) \leftarrow \mathsf{isa}(x,y), \mathsf{isa}(y,z)\}$, the atoms entailed by the KB $K_0 = (D_0, O_0)$ are:

$$ent(K_0) = D_0 \cup \{\mathsf{isa}(a,c) : \mathsf{isa}(a,b), \mathsf{isa}(b,c) \in D_0\}.$$

Consider now the set $\mathbf{U}_0 = \{\langle\mathsf{Discovery\ Cove}\rangle, \langle\mathsf{Epcot}\rangle\}$ (referred to as an *anonymous relation* or a *unit*). To comprehensively express the nexus of similarity between the elements of $\mathbf{U}_0$ and unveil its expansions, it is necessary to establish a consensus on the relevant features describing any entity in D_0. Since such features might vary depending on the specific application scenario, we introduce the notion of *summary selector*, an algorithm that, for each e in D_0, selects a subset of $ent(K_0) \cup \{\top(e) : e \text{ is an entity in } K_0\}$, referred to as the *summary* of e in the given scenario. For

[1] A first implementation of the framework is available, as a web service, at https://n9.cl/6ghqj; a systematic performance evaluation is currently underway.

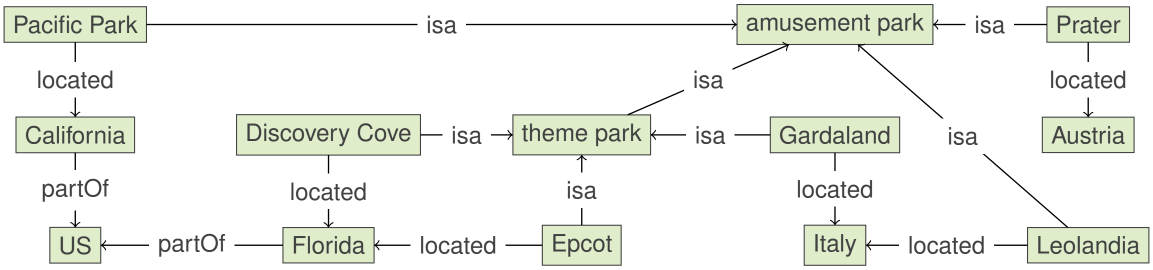
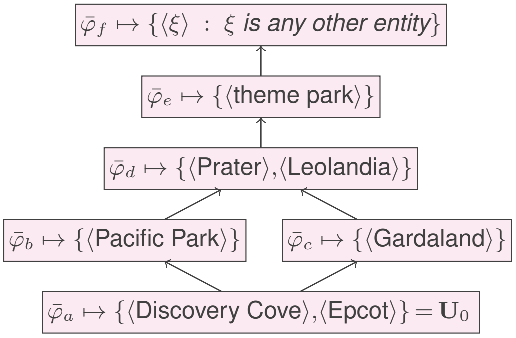


Figure 1: Knowledge graph $\mathcal{G}_0$ underlying the selective knowledge base $\mathcal{S}_0$ of our Example.


$\bar{\varphi}_f \mapsto \{\langle\xi\rangle\ :\ \xi$ *is any other entity*$\}$
$\bar{\varphi}_e \mapsto \{\langle$theme park$\rangle\}$
$\bar{\varphi}_d \mapsto \{\langle$Prater$\rangle,\langle$Leolandia$\rangle\}$
$\bar{\varphi}_b \mapsto \{\langle$Pacific Park$\rangle\}$
$\bar{\varphi}_c \mapsto \{\langle$Gardaland$\rangle\}$
$\bar{\varphi}_a \mapsto \{\langle$Discovery Cove$\rangle,\langle$Epcot$\rangle\} = \mathbf{U}_0$

$$\bar{\varphi}_f = x \leftarrow \top(x)$$

$$\bar{\varphi}_e = x \leftarrow \mathsf{isa}(x, \mathsf{ap}), \top(x), \top(\mathsf{ap})$$

$$\bar{\varphi}_d = x \leftarrow \mathsf{isa}(x, \mathsf{ap}), \mathsf{located}(x, y), \top(x), \top(y), \top(\mathsf{ap})$$

$$\bar{\varphi}_c = x \leftarrow \mathsf{isa}(x, \mathsf{tp}), \mathit{conj}(\bar{\varphi}_d), \top(\mathsf{tp})$$

$$\bar{\varphi}_b = x \leftarrow \mathsf{isa}(x, \mathsf{ap}), \mathsf{located}(x, y), \top(x), \top(y), \top(\mathsf{ap}), \mathsf{partOf}(y, \mathsf{US}), \top(\mathsf{US})$$

$$\bar{\varphi}_a = x \leftarrow \mathsf{isa}(x, \mathsf{tp}), \mathit{conj}(\bar{\varphi}_e), \mathsf{located}(x, \mathsf{Florida}), \mathsf{partOf}(\mathsf{Florida}, \mathsf{US}), \top(\mathsf{tp}), \top(\mathsf{Florida}), \top(\mathsf{US})$$

Figure 2: Expansion graph $eg(\mathbf{U}_0, \mathcal{S}_0)$, where $\mathsf{tp} = \mathsf{theme\_park}$, $\mathsf{ap} = \mathsf{amusement\_park}$, and $\mathit{conj}(\varphi)$ is the conjunction of atoms of φ.

our purposes, adopt the simple yet effective selector ς_0 that builds, for each e in D_0, the dataset $\varsigma_0(\langle e\rangle)$ as the union of:

$$\begin{aligned}
&A = \{p(f, g) \in \mathit{ent}(K_0) : f = e\},\\
&B = \{p'(f, g) \in \mathit{ent}(K_0) : p(e, f) \in A \wedge p \neq \mathsf{isa} \wedge p' \neq \mathsf{isa}\},\\
&C = \{\top(f) : f \text{ is an entity in } A \cup B\} \cup \{\top(e)\}.
\end{aligned}$$

For instance, the summary of $\langle$Epcot$\rangle$ is the dataset $\varsigma_0(\langle\mathsf{Epcot}\rangle)$ obtained by the union of:

$$\begin{aligned}
&\{\mathsf{isa}(\mathsf{Epcot}, \mathsf{tp}), \mathsf{isa}(\mathsf{Epcot}, \mathsf{ap}), \mathsf{located}(\mathsf{Epcot}, \mathsf{Florida})\},\\
&\{\mathsf{partOf}(\mathsf{Florida}, \mathsf{US})\},\\
&\{\top(\mathsf{Epcot}), \top(\mathsf{tp}), \top(\mathsf{ap}), \top(\mathsf{Florida}), \top(\mathsf{US})\}.
\end{aligned}$$

We refer to the pair $\mathcal{S}_0 = (K_0, \varsigma_0)$ as a *selective KB*. By examining the formula

$$\bar{\varphi}_1 = x \leftarrow \mathsf{isa}(x, \mathsf{ap}), \mathsf{located}(x, y), \mathsf{partOf}(y, \mathsf{US})$$

in relation to the considered summaries, it is evident that $\bar{\varphi}_1$ *explains* some nexus of similarity between $\langle$Discovery Cove$\rangle$ and $\langle$Epcot$\rangle$. However, $\bar{\varphi}_1$ neglects the additional information that both entities are also located in Florida according to their summaries. Conversely, the formula $\bar{\varphi}_a$ in Figure 2 fully explains the nexus of similarity between the two entities. Hence, we can assert that $\bar{\varphi}_a$ *characterizes* their nexus of similarity.

The last step is to classify each entity e of $\mathcal{S}_0$ in relation to $\mathbf{U}_0$, by characterizing each unit $\mathbf{U}_0 \cup \{\langle e\rangle\}$. This leads to the *expansion graph* of $\mathbf{U}_0$ with respect to $\mathcal{S}_0$, denoted by $eg(\mathbf{U}_0, \mathcal{S}_0)$ and depicted in Figure 2. Intuitively, each node n_1 labeled by $\varphi_1 \mapsto \mathbf{U}_1$ says that φ_1 characterizes $\mathbf{U}_0 \cup \{\langle e\rangle\}$ for each $\langle e\rangle \in \mathbf{U}_1$. If there is a path from n_1 to another node n_2 labeled by $\varphi_2 \mapsto \mathbf{U}_2$, it means that φ_2 characterizes the unit $\mathbf{U}_0 \cup \mathbf{U}_1 \cup \mathbf{U}_2$ as well. Thus, we can conclude, for instance, that the nexus of similarity that $\langle$Gardaland$\rangle$ has with $\mathbf{U}_0$ incorporate those that $\langle$Leolandia$\rangle$ has with $\mathbf{U}_0$, showing that Gardaland is more similar to the entities of $\mathbf{U}_0$ than Leolandia with respect to $\mathcal{S}$. Additionally, the nexus of similarity that $\langle$Pacific Park$\rangle$ has with $\mathbf{U}_0$ are incomparable to those that $\langle$Gardaland$\rangle$ has with $\mathbf{U}_0$. In simple terms, $eg(\mathbf{U}_0, \mathcal{S}_0)$ is the expected taxonomic expansion of $\mathbf{U}_0$.

Acknowledgments. This work was funded by: FAIR (CUP: PE00000013) Spoke 9 and Tech4You (CUP: ECS0000009) Spoke 6, under the NextGenerationEU NRRP MUR program; and PRODE (CUP: H53D23003420006) under the PRIN MUR program.

References

Amendola, G.; Manna, M.; and Ricioppo, A. 2024. A logic-based framework for characterizing nexus of similarity within knowledge bases. *Information Sciences* 664.

Cirasella, J. 2007. Google sets, google suggest, and google search history: Three more tools for the reference librarians bag of tricks. *The Reference Librarian* 48(1).

Cohen, W. W.; Borgida, A.; and Hirsh, H. 1992. Computing least common subsumers in description logics. In *AAAI'92*.

Colucci, S.; Donini, F. M.; Giannini, S.; and Sciascio, E. D. 2016. Defining and computing least common subsumers in RDF. *J. Web Semant.* 39.

Gomaa, W., and Fahmy, A. 2013. A survey of text similarity approaches. *Int. J. Comput. Appl.* 68(13).

Petrova, A.; Kostylev, E. V.; Grau, B. C.; and Horrocks, I. 2019. Query-based entity comparison in knowledge graphs revisited. In *ISWC'19*.

ten Cate, B.; Dalmau, V.; Funk, M.; and Lutz, C. 2023. Extremal fitting problems for conjunctive queries. In *PODS'23*.