A Strategy for Analyzing and Enriching Domain Ontologies

Mariela Rico[1], Ma. Laura Caliusco[1], Emiliano Reynares[1],
Omar Chiotti[1,2], and Ma. Rosa Galli[1,2]

[1] CIDISI Research Center, UTN-FRSF, Lavaise 610, S3004EWB Santa Fe, Argentina
{mrico,mlcaliusco}@santafe-conicet.gov.ar,reynares.emiliano@gmail.com
[2] INGAR-UTN-CONICET, Avellaneda 3657, S3002GJC Santa Fe, Argentina
{chiotti,mrgalli}@santafe-conicet.gov.ar

Abstract. During the last years there is an increasing interest on the ontologies. The way in which the semantics of a domain entity is represented in an ontology may differ depending on the requirements that the ontology should satisfy. However, in general the richness is a desired quality attribute for an ontology.
The contribution of this paper is a strategy for evaluating and enriching the representation of real entity semantics in a domain ontology. In addition, an application example is discussed.

1 Introduction

During the last years, with the advent of the Semantic Web, there is an increasing interest on the ontologies [1]. The way in which the semantics of a domain entity is represented in an ontology may differ depending on the requirements that the ontology should satisfy [2]. However, in general the richness is a desired quality attribute for an ontology.

When representing the semantics of an entity in an ontology, it is possible that some of its features are implicit or the representation of these features is incomplete. Therefore, their explicitness becomes necessary for enriching the representation of the entity semantics. In this way, a strategy is necessary to evaluate and improve the representation of the entity semantics in an ontology.

The purpose of this paper is to present a strategy for analyzing and enriching the representation of the semantics of domain entities in an existing domain ontology. To this aim, Section 2 defines the main concepts of the paper. Section 3 describes the proposed strategy. Section 4 shows an application example. Finally, Section 5 is devoted to the conclusions and future work.

2 Background

The objective of this section is to give definitions to a set of main terms, which will be used in the remainder of the paper.

A **domain** is a portion of reality that forms the subject-matter of a single field of study, a technology, or a mode of study; i.e., the domain of computer science, of e-commerce, among others. A **domain entity** is anything which exists within a given domain, including objects, processes, qualities and states [3].

A **context** is a set of circumstances in which something exists or occurs. In a context, it can be distinguished features that describe that circumstances such as geopolitical conditions, factors influencing a process, and so on. A given domain entity or its features can have different interpretations and/or representations depending on the context in which it is considered.

A **contextual feature** is one whose interpretation and / or representation depend on the features of the context in which the domain entity is considered.

The semantics of domain entities could be represented by a **domain ontology** (from now on "ontology") that is a representational artifact used to render the cognitive representations of a given domain [3]. The ontology elements are: **Terms** that are words or group of words that represent domain entities in a given domain or entity features that can be considered entities in themselves; **Properties** that represent the features of domain entities that have not been considered entities in themselves; **Relations** that are elements that glue together other ontology elements; and **Axioms** that serve to represent sentences that are always true in a domain [1].

3 A Strategy for Enriching Domain Ontologies

The strategy proposed in this paper intends to improve the representation of domain entity semantics by representing their contextual features. It is composed by six steps shown in Figure 1.

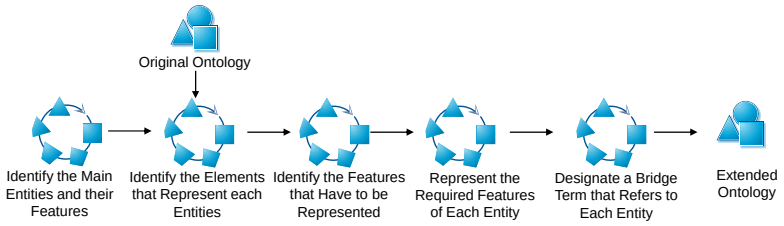

Fig. 1. A Strategy for Enriching Domain Ontologies by Representing Contextual Features.

3.1 Step 1: Identify the Main Entities and their Features

The aim of this step is to specify what the main domain entities should be represented in the ontology, their features, and their relations with other entities.

The outputs of this step are both a list of the entities and their relations, and, for each entity, a list of their features, whose semantics is affected by the context in which the entity is considered. The list of the entities can be obtained by exploring the documentation associated with the domain of discourse or by having meetings with domain experts. The list of the features can be made by encoding the knowledge required to understand the corresponding entities. If necessary, domain experts could help to identify the features by indicating the underlying knowledge they assume should be possessed in order to have an accurate interpretation of the entity meaning in different contexts.

3.2 Step 2: Identify the Elements that Represent each Entity

The objective of this step is to identify the ontology elements that represent each entity and their features as recognized in the Step 1. An entity could have been represented with only a simple term or with a set of ontology elements.

3.3 Step 3: Identify the Features that Have to be Represented

This step allows the ontologist to identify the features, which are generally implicit in the representation of an entity. These features are not required to be made explicit when the entities are considered within the same context, but it becomes necessary when the entities have to be interpreted in another one. Based on the outputs of previous steps, these features can be detected. Since not all of these features need to be made explicit, answering the following questions could help to identify the ones that do need.

- Are there any implicit features in the representation of an entity that although they may be inferred by a human they cannot be inferred by a machine? If the answer is yes, could these features be inferred in the wrong way in other contexts different from the considered one? If the answer is yes, these features should be represented.
- Are there any entities whose representations and/or meanings could change depending on the context in which the entities are considered? If the answer is yes, are the representations and meanings of the features or entities completely explicit in the ontology? If the answer is no, the representations and meanings should be made explicit.
- What are the dimensions used to represent a feature? Are they the same regardless of the context in which the feature is considered? If the answer is no, are they explicit in the ontology? If the answer is no, these dimensions should be represented.

3.4 Step 4: Represent the Required Features of Each Entity

Once the features and their dimensions have been identified, they have to be represented in the ontology. For that, reusing existing ontologies should be considered. This can be made by importing the portion of the ontology to be reused.

When it is not possible to reuse an ontology, it must to identify if the feature is simple or complex. A simple feature is a quality that does not bear other qualities, and it is associated with a one-dimensional representation in human cognition [4]. For example, the weight of a thing is associated with a one-dimensional structure, whose possible values are positive real numbers. Thus, two elements should be added to the ontology: a term denoting the representation dimension, and a relation between this term and the term that represents the simple feature.

A complex feature is a quality that bears other qualities, and it is associated with a set of integral dimensions that are separable from all other dimensions [4]. An integral dimension is one in which it is not possible to assign a value to an object on one dimension without giving it a value on the other. For instance, color can be represented in terms of the dimensions of hue, saturation, and brightness. These dimensions are integral. By contrast, weight and hue dimensions are said to be separable. Each integral dimension is associated with a simple feature. In order to improve the representation of a complex feature, the following elements should be added to the ontology:

- A term representing the set of integral dimensions and a relation between this term and the term that represents the complex feature.
- For each integral dimension, a term representing it and a relation between this term and the term that represents the set of integral dimensions.
- For each term representing an integral dimension, a relation between this term and the term that represents the corresponding simple feature.

In addition, for each term representing a one-dimensional representation or an integral dimension, a term representing the unit of measurement of the dimension and a relation between these two terms should be added to the ontology.

Finally, it is possible that the feature to be represented is modeled as an attribute of a term. In this case, this attribute should be deleted. It is important to denote that in this way an attribute will become a term, and the ontology will be more expressive due to the fact that the reasoners work better.

3.5 Step 5: Designate a Bridge Term that Refers to Each Entity

The intended uses of an entity in a given context should be represented by a term, called bridge term because it allows linking different meanings and representations of the same entity in different contexts. A bridge term should also be interpreted as representing contextual features, because the intended use depends on the context in which the entity is considered. Thus, in the ontology, it is necessary to determine if there is a term that designates the intended use of each entity in the considered context, otherwise, it has to be added.

A bridge term should be related to the elements that represent the entity whose intended use it represents. An entity could be represented by a single element or a set of them. In the former case, a relation between that single element and the bridge term should be added. In the latter case, the most representative term should be chosen and then, a relation between this term and the bridge

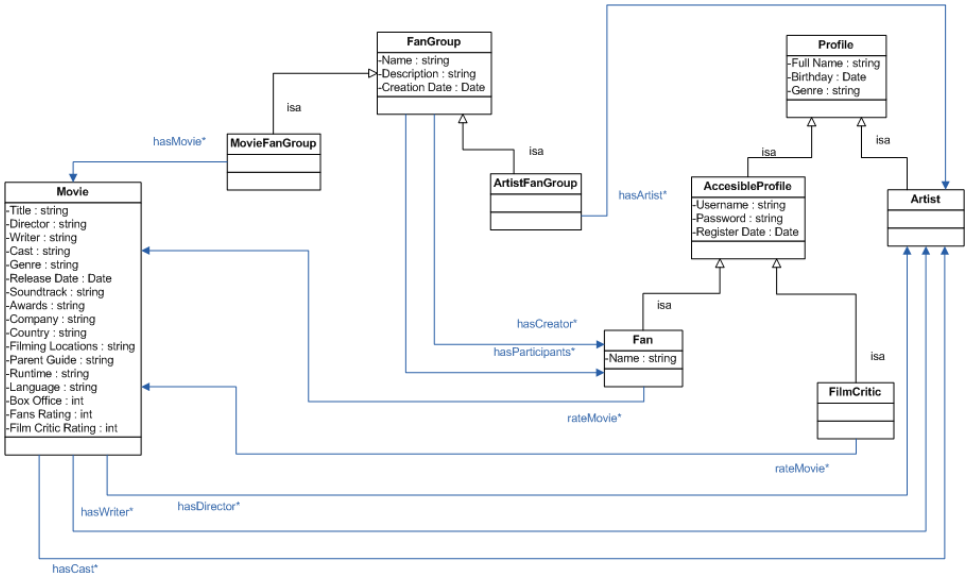

Fig. 2. Original SOCIAL MOVIE SITE ontology (O_{SMS})

term should be added. As a bridge term represents a contextual feature, it should also be related to the term that represents its dimension.

4 An Application Example

A good ranking for a website in a semantic search engine depends on the quality of the ontology that represents the meaning of the website. Generally, the process of improving ranking of a website in a search engine is an iterative and human-oriented process.

The strategy proposed in this paper could be applied with the aim of supporting the process of improving ranking of a website in a semantic search engine by enriching the ontology that represents its content. Figure 2 shows an ontology of a social network to share tastes in movies. Following, the steps to enrich the SOCIAL MOVIE SITE ontology (O_{SMS}) is presented.

4.1 Enriching the Ontology

Step 1: Identify the Main Entities and their Features The main entities involved in the social network site are:

– A **movie** is the main entity of the system. Its features are a title, director, writer, cast, genre, release date, soundtrack, awards, among others. The attributes of interest from the social network viewpoint are the scores given by film critics and fans.

– A **fan** is a person with an intense linking and enthusiasm for a movie. A fan usually give a score to a movie.
– A **film critic** is a person who gives an analysis and evaluation of films.

Step 2: Identify the Ontology Elements that Represent Each Entity A movie is represented by the term `Movie`, their properties and the relations `hasDirector`, `hasWriter` and `hasCast` between this term and the term `Artist`. An artist participating in any of the movies is represented by the terms `Profile` and `Artist`, their properties and the relation `is-a` between them.

A fan is represented by the terms `Profile`, `AccesibleProfile` and `Fan`, their properties, the relation `is-a` between these terms and the relation `rateMovie` with the movies the fan has assigned a score. A group of fans of a given movie is represented by the terms `FanGroup` and `MovieFanGroup`, their properties, the relation `is-a` between these terms, and the relations `hasCreator`, `hasMovie` and `hasParticipants`. A group of fans of a given artist is represented by the terms `FanGroup` and `ArtistFanGroup`, their properties, the relation `is-a` between these terms, and the relations `hasCreator`, `hasArtist` and `hasParticipants`.

A film critic is represented by the terms `Profile`, `AccesibleProfile` and `FilmCritic`, their properties, the relation `is-a` between these terms, and the relation `rateMovie` with the movies that the critic has assigned a score.

Step 3: Identify the Features that Have to be Represented Since a movie is the main entity in the social network and their main attribute are the scores of the fans and film critics, we have to pay attention in their representation. In the case of fans, they qualify the movies by an integer value between 0 and 10. The film critics in general consider different aspects to give a qualification to a certain movie. However, in the original O_{SMS}, they only can give an integer value between 0 and 10. Making explicit the aspects considering by film critics will improve the representation of the qualification concept.

The genre, the parent guide, and duration of a movie are represented by a string. This representation does not satisfy the minimal encoding bias criterion [5]. The box office of a movie is represented by an integer. Then, the following questions arise: What is the unit of measure? Dollars or Euros? The value is expressed in miles or millions? These issues must be solved in order to avoid misunderstandings.

Finally, a movie has an attribute that represents its release date, the profiles have an attribute that represent the date of birth and, the fans have a creation date. All of these attributes are represented by a datatype date not satisfying the minimal encoding bias criterion.

Step 4: Represent the Required Features of Each Entity With the aim of improving the representation of a particular date in the O_{SMS}, reusing the OWL-Time ontology[3] was considered. The release date of a movie is represented

[3] http://www.w3.org/TR/owl-time/

by the term `ReleaseDate`, that derives from the term `CalendarInstant`, and is related to the term `Movie` by using a formal relation (See Fig. 3).

Since the box office of a movie is a simple feature, six elements should be added to the O_{SMS}: a term `BoxOfficeDimension` denoting the representation dimension, a term `UnitOfMeasure` representing the unit of measure of the dimension, a relation between these terms, and a relation between the terms `BoxOfficeDimension` and `BoxOffice`. Finally, this term has to be related with the term `Movie`. An equivalent procedure has to be carried out considering the simple feature that represents the movie duration. In this case, it has to be added the term `RuntimeDimension` related to the term `UnitOfMeasure`, and the relation between the terms `RuntimeDimension` and `Runtime` and this term with the term `Movie` (See Fig. 3).

With respect the fan score, the process presents a unique difference. Applicable scores include integers between 0 and 10, but do not have a unit of measurement. Simply, the higher the score more like a film. In this way the term `FanRatingDimension` and a relation with the term `FanRating` are added. Finally, the term `FanRating` is associated with the term `Movie`.

With the aim of representing the genre and the parental guidance, the terms `Genre` and `ParentGuide` are added to the O_{SMS}. Each of them is related to the terms `GenreDimension` and `ParentGuideDimension`. However, these dimensions are not quantitative but consist of a list of possible values. For example, `GenreDimension` could consists of a enumeration like action, adventure, drama, comedy, romance, thriller, while the ParentGuideDimension could consist of G, PG-13, PG-18, PG-21(See Fig. 3).

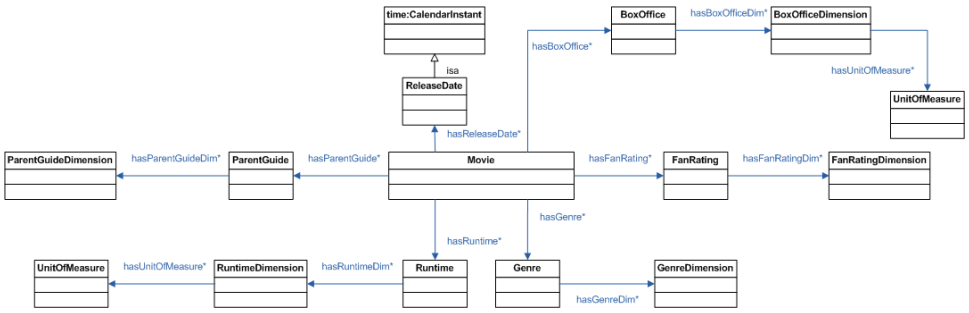

Fig. 3. A portion of the extended O_{SMS} representing fan rating, run time and parent guidance features.

Finally, representing the criteria generally used to describe a film critics would allow a more complete representation of this measure. In general, an evaluation is done based on three basic aspects of a movie: The story told by the film, the quality of the picture and the quality of soundtrack. To represent this, three terms are added: `StoryRating`, `PhotographRating` and `SoundtrackRating`. These

terms have a formal relation with the term `Movie`. Their dimensions are represented by the terms `StoryRatingDimension`, `PhotographRatingDimension` and `SoundtrackRatingDimension`, respectively. The unit of measure is not considered, since evaluations consist of integer values between 0 and 10. However, the evaluation includes these terms and is a complex feature related to a set of dimensions represented by the term `ReviewerRatingMultidimension`. The dimensions of this set are integral dimensions represented by the terms `StoryRatingDimension`, `PhotographRatingDimension` and `SoundtrackRatingDimension`. Figure 4 shows a portion of the extended O_{SMS} representing the aforementioned features.

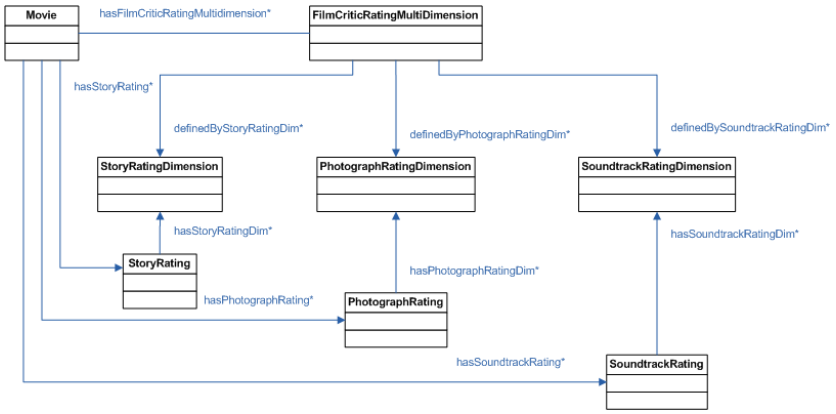

Fig. 4. A portion of the extended O_{SMS} representing the movie rating of critics.

Step 5: Designate a Bridge Term that Refers to Each Entity Most of the modelling concepts seems to be clear in the context of the social networks considered. However, the term `AccesibleProfile` could cause a misunderstanding. With the aim of clarifying its intended use and meaning, the bridge term `User` is added. This term represents the meaning of the agreed in the context of social networks a those who are able to access the system through a username and password. Figure 5 shows the extended O_{SMS} representing the bridge term `User`.

4.2 Evaluating the Ontology Enrichment

The application of the strategy have generated a rich extended O_{SMS} that represents a more detailed knowledge in a coherent and compact manner than the original O_{SMS}. Then, the extended O_{SMS} should have a better ranking position. This can be demonstrated by applying the metrics defined by Alani et. al [6] as shown in Table 1.

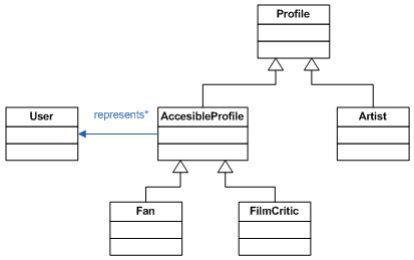

Fig. 5. A portion of the extended O_{SMS} showing the bridge term `User`

Table 1. Evaluation of ontology representation improvement

Metrics	Original O_{SMS}	Extended O_{SMS}	Improvement
Class Match Measure (CMM)	3,4	4,6	35%
Density Measure (DEM)	6	7,4	23%
Semantic Similarity Measure (SSM)	5	9,3	46%

– Class Match Measure (CMM). It is meant to evaluate the coverage of an ontology for a given term.
Let C be a set of classes in an ontology, and T is the set of main entities identified in Step 1. Each entity will be representated by its main label, i.e. `Movie`, `Fan` and `Film Critic`.

$$E = \sum_{c \in C} \sum_{t \in T} I(c,t) \text{ where } \begin{array}{l} I(c,t) = 1 : \text{if } label(c) = t \\ I(c,t) = 0 : \text{if } label(c) \neq t \end{array}$$

$$P = \sum_{c \in C} \sum_{t \in T} J(c,t) \text{ where } \begin{array}{l} J(c,t) = 1 : \text{if label(c) contains t} \\ J(c,t) = 0 : \text{if label(c) not contain t} \end{array}$$

$$CMM = 0.6 * E + 0.4 * P$$

– Density Measure (DEM). It is intended to approximate the representational-density or information-content of classes that matches the main entities and consequently the level of knowledge detail. An ontology that contains many relations other than class-subclass relations is richer than a taxonomy[7].

$$DEM = \\ (\tfrac{1}{n}) \sum_{c \in C} (2 * relations[c] + superclasses[c] + subclasses[c] + siblings[c])$$

where n is the number of matched classes.

– Semantic Similarity Measure (SSM). It calculates how close the classes that matches the main entities are in an ontology. SSM is measured from the minimum number of relations that connects a pair of terms.

Let c_i, $c_j \in C$ and $c_i \rightarrow^p c_j$ is a path $p \in P$ of paths between c_i and c_j.

$$ssm(c_i, c_j) = \begin{cases} \frac{1}{length(min_{p \in P\, c_i \rightarrow^p c_j})} & : \text{if } i \neq j \\ 0 & : \text{if } i = j \end{cases}$$

$$SSM = (\tfrac{1}{n}) \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} ssm(c_i, c_j)$$

5 Conclusions and Future Work

In this paper, a strategy for evaluating and enriching the semantic representation of domain entities in an ontology was presented. This strategy could be applied in different scenarios. In this paper, the strategy was applied to enrich an ontology with the aim of improving the ranking of the website that it represents.

A complementary approach for increasing the expressiveness of an ontology is the addition of rules. Future work will be focused on defining a methodology that includes the definition and implementation of rules.

References

1. Gómez-Pérez, A., Fernández-López, M., Corcho, O.: Ontological engineering. 2nd edn. Springer Verlag, New York (2004)
2. Gangemi, A., Catenacci, C., Ciaramita, M., Lehmann, J.: Modelling Ontology Evaluation and Validation. In: ESWC. (2006) 140–154
3. Smith, B., Kusnierczyk, W., Schober, D., Ceusters, W.: Towards a reference terminology for ontology research and development in the biomedical domain. In: 2nd Int. Workshop on Formal Biomedical Knowledge Representation. (2006) 57–66
4. Guizzardi, G.: Ontological foundations for structural conceptual models. PhD thesis, University of Twente, Enschede, The Netherlands (2005)
5. Grüber, T.: A translation approach to portable ontology specification. Knowl. Ac. **5**(2) (1993) 199–220
6. Alani, H., Brewster, C.: Metrics for ranking ontologies. In: 4th Int. EON Workshop, 15th Int. World Wide Web Conf. (2006)
7. Tartir, S., Arpinar, I.B., Moore, M., Sheth, A.P., Aleman-meza, B.: Ontoqa: Metric-based ontology quality analysis. In: Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources. (2005)