Extending Semantic Similarity Measurement with Thematic Roles

Krzysztof Janowicz

Institute for Geoinformatics,
University of Muenster, Germany
janowicz@uni-muenster.de

Abstract. Semantic similarity measurement plays a significant role in semantic interoperability and in information retrieval within the geo domain as it supports the detection of conceptually close but not identical entities. In feature-based models, the similarity measurement is done by comparing common and different features such as parts, attributes and functions. This paper suggests adding thematic roles as an additional type of features to be compared, and shows why and how the usage of thematic roles may prevent wrong function matches.

1 Introduction

Ontologies specify a conceptualization of entities represented in geographic information systems (and services), and therefore allow the users to interpret the meaning of the used terms. What makes information retrieval and usage difficult is that users often have no clear class (concept) definition in their mind that could be compared to the specification of the geographic information system or both definitions do not match. Semantic similarity measurement offers the possibility to define an area of interest and to calculate the distance between the classes within this area. In contrast to rigid logic-based reasoning, the result should be more flexible and adaptable, and therefore close the gap between user-expected and system-retrieved meanings.

The Matching-Distance Similarity Measure (MDSM) [1] is such a (feature-based) measurement theory introduced for the geo domain. The intension of this paper is to present an extension to MDSM that is able to measure similarity based on the idea that entity classes whose members share a certain behavior are similar. Thematic Roles are used to model this behavioral aspect, because they offer an abstract theory (that is grounded in Sowa's [2] formal ontology) of roles an entity plays within a certain function.

The goal of this extension is to avoid wrong matches within the functional feature (FF) similarity calculation of MDSM and to improve the robustness of the model by aligning the entity classes to roles described within formal ontology.

M.A. Rodríguez et al. (Eds.): GeoS 2005, LNCS 3799, pp. 137 – 152, 2005.

2 Related Work

This section introduces Thematic Roles (TR), Matching-Distance Similarity Measure, Role-Governed and Transformational Categories as foundation for the semantic similarity measurement extension presented in this paper.

2.1 Thematic Roles

Influenced by the work of Moravcsik [3], Dick [4] and Pustejovsky [5], John Sowa [6] related Somers [7] case grid to Aristotle's idea of four causes (efficient cause, material cause, final cause and formal cause) called *aitiai*. The result is a matrix of six rows representing verb categories (or to be more precise the type of nexus [6]) and four columns representing different kinds of participants. Each of the twenty-four cells represents at least one thematic role such as Agent or Location. These thematic roles are arranged within a hierarchy of participants depending on their position in the matrix. At the top of this hierarchy Source and Product participants are distinguished. At the next level Source is further distinguished into the Initiator and Resource participants and Product subsumes the Goal and Essence participants. Location for example is a special kind of Essence (Location<Essence<Product<Participant) (see Figure 1). In contrast to roles in description logics thematic roles are not binary relations but concepts (unary predicates) [8].

With respect to entities this means that, depending on the context, each entity plays a specific thematic role. For example, a Person who arrives at a sport arena is

	Source		Product	
	Initiator	Resource	Goal	Essence
Action	Actor, Effector	Instrument	Result, Recipient	Patient, Theme
Process	Agent, Origin	Matter	Result, Recipient	Patient, Theme
Transfer	Agent, Origin	Instrument, Medium	Experiencer, Recipient	Theme
Spatial	Origin	Path	Destination	Location
Temporal	Start	Duration	Completion	Point in Time
Ambient	Origin	Instrument, Matter	Result	Theme

Fig. 1. Thematic roles matrix [2][6]

regarded as Actor whereas the sport arena is regarded as Location. The corresponding conceptual graph representation looks as follows:

[Person: Bob]←(Agnt)←[Arrive]→(Loc)→[Sport Arena]

Due to the hierarchical structure of participants the conceptual graph can be shifted on a more abstract level. This is very useful in case of ambiguity, i.e. when it is not clear what role is played by a certain entity [2][6], or for comparison between different cases as discussed later in this paper.

[Person: Sue]←(Agnt)←[Go]→(Dest)→[City: Mexico City]

In both cases the Person plays the role of an Agent and therefore no shift to Initiator is necessary whereas Location (Location<Essence<Product) and Destination (Destination<Goal<Product) have to be replaced by Product which is their immediate common superclass. Entities are not restricted to occupy the same thematic role in different cases and therefore Bob becomes the Recipient (Goal) in "*Sue sent the gift to Bob by Federal Express*" [2, p. 506].

Sowa [2][6] places the thematic roles in an intermediate level of his formal ontology and suggests creating subtypes for each kind that is of interest for a certain domain or context (e.g. TaxiDriver<Driver<Doer<Agent<Initiator<…). Sowa argues that Driver only represents persons who are actively driving a vehicle and that therefore a LicensedDriver (e.g. Chauffeur) can not be a subtype of Driver e.g. because licensed drivers are legally authorized to drive a vehicle whether they are driving it right now or not.

2.2 Matching-Distance Similarity Measure

MDSM is the asymmetric and context sensitive semantic similarity measurement approach for entity classes developed by Rodriguez and Egenhofer [1]. It can be regarded as an extension of Tverskys [9] ratio model and therefore is classified as a feature-based approach to similarity (in contrast to geometric and alignment models for example [10][11][12] which calculates the similarity using the number of common and different features. Three kinds of features can be distinguished: parts, which are structural components of a class such as wall for building; functions which describe "what is done to or with a class" [1, p. 232} such as the function educate is offered by college (the idea of functions in MDSM is close to Gibson's [13] affordances) and attributes which are additional characteristics that can not be regarded as parts or functions such as name or owner type for building.

$$S(c_1,c_2) = \omega_p \cdot S_p(c_1,c_2) + \omega_f \cdot S_f(c_1,c_2) + \omega_a \cdot S_a(c_1,c_2) \tag{1}$$

Equation 1 displays the overall semantic similarity measurement, which is regarded as the sum of the weighted similarities of the three kinds of features (parts, attributes and functions) of the compared entity classes c1 and c2.

$$S_t(c_1,c_2) = \frac{\left|C_1 \cap C_2\right|}{\left|C_1 \cap C_2\right| + \alpha(c_1,c_2) \cdot \left|C_1 \setminus C_2\right| + (1-\alpha(c_1,c_2)) \cdot \left|C_2 \setminus C_1\right|} \tag{2}$$

Equation 2 describes the no-symmetric similarity function for each of the feature types. S_t (c1, c2) is defined as the similarity for the feature type t between the entity classes c1 and c2 where C_1 and C_2 are the sets of features of type t for c1 and c2, $|C_1 \cap C_2|$ is the cardinality of the set intersection and $|C_1 \setminus C_2|$ is the cardinality of the set difference.

The relative importance α (equation 3) of the different features of type t is defined in terms of the distance d between c1 and c2 within a hierarchy that takes taxonomic and partonomic relations into account. Lub denotes the least upper bound, i.e. the immediate common superclass of c1 and c2 [1].

$$\alpha(c1,c2) = \begin{cases} \dfrac{d(c1,\text{lub})}{d(c1,c2)}, d(c1,\text{lub}) \le d(c2,\text{lub}) \\ 1 - \dfrac{d(c1,\text{lub})}{d(c1,c2)}, d(c1,\text{lub}) > d(c2,\text{lub}) \end{cases} \quad where: d(c1,c2) = d(c1,\text{lub}) + d(c2,\text{lub}) \tag{3}$$

MDSM takes context into account and therefore the weighting in the overall similarity function (equation 1) is calculated depending on the domain of application using variability or commonality within the features (of each type). "Contextual information (C) is specified as a set of tuples over operations (op_i) associated with their respective noun arguments (e_j) (Equation 4). The nouns correspond to entity classes in MDSM, while the operations refer to verbs that are associated with methods of these classes." [1, p. 239] A contextual specification such as C = <(play, {})> for example expresses a domain of application that contains all entity classes which share the functional feature play.

$$C = \left\langle (op_1, \{e_1, ..., e_m\}), ..., (op_n, (e_1, ..., e_l)) \right\rangle \tag{4}$$

Within such a context the relevance (ω_t in equation 1) of each feature type is defined either by the variability P_t^v (equation 5) or commonality P_t^c function (equation 6) and then normalized with respect to the remaining feature types so that $\omega_p+\omega_f+\omega_a$ is always 1.

$$P_t^v = 1 - \sum_{i=1}^{l} \frac{o_i}{n \cdot l} \tag{5}$$

The variability describes how diagnostic [9][14] a feature type t is within a certain domain of application by assuming that the more characteristic each feature is for a given class the more diagnostic it is. A certain feature of type t has low relevance if it appears in many classes and high relevance if it is not common to the classes within the domain. P_t^v is the sum of the diagnosticity of all features of the type t in the domain and therefore 0 when all features a shared by all entity classes (P_t^v=1-1=0) and close to 1 if each feature is unique (where o_i is the number of occurrences of the feature within the domain) and the number of features l and classes n in the domain is high.

$$P_t^c = \sum_{i=1}^{l} \frac{o_i}{n \cdot l} = 1 - P_t^v \tag{6}$$

Commonality is defined as the opposite of variability ($P_t^c = 1 - P_t^v$) and assumes that by defining a domain of application the user implicitly states what features are relevant [1].

2.3 Role-Governed and Transformational Categories

Depending on the classification and the level of granularity several kinds of categories can be distinguished. Beside common also called feature-based categories, role-governed and transformational categories are of special interest for this paper.

Role-Governed Categories
In contrast to common categories, members of role-governed categories are not grouped together because they share a set of necessary (and sufficient) features, but due to a certain role they play within a domain or context [15][16]. Wittgenstein [17] argued, that it is difficult to find a feature-based representation of Game, but as described by Markman and Stilwell [15] Game may be regarded as a role governed category that is specified as being the second argument of the relation play(Player, Game) where Player is also defined in a role-governed way. In other words, games are the entities played by players.

The coherence between the category members is merely based on few (or even one) significant core roles and therefore the overall similarity between the members is low in general [16][18]. Nevertheless graded structure also exists for non feature-based categories and therefore similarity measurement is possible in principle [19].

Moreover role-governed categories cannot be arranged within feature hierarchies as this is possible for feature-based categories. On the one hand they do not inherit properties and on the other hand - besides a very abstract functional theory - they do not necessarily share a common role [15].

The importance of social roles for concepts such as money or president is discussed by Masolo et al. [8]. The importance of roles in the geo domain is elucidated by Kuhn [20].

Transformational Categories
Markman and Stilwell [15] claim that there is an additional kind of categories called transformational categories that specify a change in a certain selection restriction for a relation. For example, according to the specification of Markman and Stilwell a player in the relation play (Player, Game) has to be a sentient being, but a team can also play a game. Thus, team is a transformational category that transforms a group to an individual. Metonymy can be regarded as a linguistic and cognitive device for the creation of transformational categories [15].

3 Why Play and Play Do Not Match

This section discusses the relation between the functions as defined in the lightweight ontology used within MDSM on the one hand and the thematic roles on the other hand.

3.1 Two Shortcomings of Feature-Based Similarity Models

Beside others [11], there are two main shortcomings that (more or less) affect all kinds of feature-based models. All features are unary, which means that an entity which is green for example, is described by the feature green and not by a feature-value pair such as color = green. On the level of entity classes an adult (in Germany) would have to be defined by the feature over17 and not age > 17. This simplification may lead to difficulties [11] for example if it is not clear whether the feature Height of the entity class Theater [1] corresponds to the height of the building or the height of the stage. The use of two separate features such as BuildingHeight and StageHeight is impractical because it will decrease similarity to all other buildings. In the case of functional features the play function of Sport arena and Game are regarded as a common feature of both classes and therefore match. This is possible because the relation between function and entity class is very loosely defined in (the lightweight ontology used in) MDSM. A function of a class can be anything that is afforded by this class independently of which role it plays within this function (and due to polysemy of verbs one could imagine many different play functions). In the upper case the entity class is either the location where one can play or the thing that is played. The KIF like code fragments below shows two simplified specifications of play, whereas the first function might also be named played-at.

```
1. (DEFRELATION PLAY (?X ?Y)
   :=> (AND (GAME ?X) (SPORTARENA ?Y)))

2. (DEFRELATION PLAY (?X ?Y)
  :=> (AND (PLAYER ?X) (GAME ?Y)))
```

MDSM is able to deal with polysemy of entity class names using taxonomic and partonomic relations but not with polysemous feature names [1].

A second weakness of feature-based similarity models (and also geometric approaches) is that they regard classes as bags of unsorted features, which means that there is no structure connecting the features within a class or even to other classes. The topological relation above(Circle, Triangle) [11] does not describe the same fact as above(Triangle, Circle). In a similarity assessment subjects may judge above(Triangle, Circle) to be more similar to above(Rectangle, Circle) than to above(Circle, Triangle) because of the same role (being under something else) that the circle plays within the first two examples (see also [21]). A first step to solve these problems is to structure the features into types as done by MDSM. Moreover Rodriguez and Egenhofer [1] propose to investigate the semantic comparison of distinguishing features (in contrast to the comparison of their labels as done so far).

As argued by Goldstone and Son [11] in extreme cases the combination of both shortcomings may lead to a spurious match in the feature-based similarity model because "a car with a green wheel and a truck with a green hood both share the feature green" [11, p. 15]. In fact the wheel and the hood share the feature green but not the compared car and truck and therefore this kind of match should not increase similarity.

3.2 Functional Features and Thematic Roles

Functions in the feature-based approach are just synsets such as {play} or {recreate, play} [1] whereas thematic roles can be regarded as entity classes which are subclasses of Participant [2]. The relation between both is that the entity class to which the functional feature belongs is a special kind of participant and the kind of function determinates its possible role. The same synset {play} may represent functions that involve different participants. These participants such as player, game or sport arena are subtypes of thematic roles. In play(Player, Game) Player is a subtype of the thematic role Agent and Game of Theme whereas in a case such as play(Player, Sport arena) the second parameter is not a subrole of Theme but of Location (see Figure 3). If both play functions are features of two different entity classes they should not be regarded as match (a feature that is common to both classes) and thus increase similarity between the compared entity classes as done in the MDSM model so far. Functional features should only count as common if the compared entity classes both occupy the same functional roles within these functions.

Sometimes it may not be clear which thematic role has to be taken out of the twenty-four cell matrix to describe the role of an entity class within a functional feature. In many cases thematic roles can be directly excluded considering their conceptual relation as described by Sowa [2]. For example, sport arena can not be a Recipient because the corresponding conceptual relation Rcpt(Act, Animate) restricts the usage of Recipient to Act and Animate. In the case of a valid allocation such as play(Agent, Location) for play(Player, Sport arena) as functional feature of the entity class Player, the resulting conceptual graph is:

`[Player]←(Agnt)←[Play]→(Loc)→[Sport Arena]`

In other cases of uncertainty it is possible to use the immediate superclass instead of a concrete functional role such as Essence for Location.

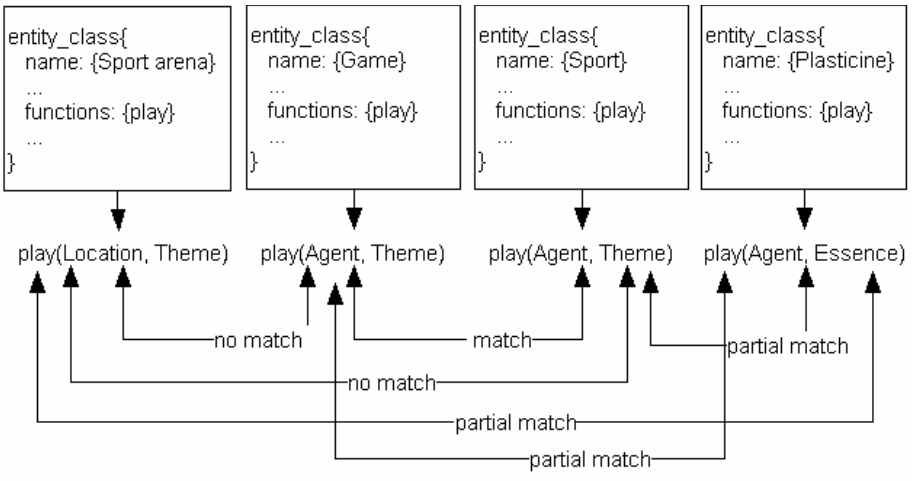

Fig. 2. Full and partial matches between the functional feature play

As depicted in Figure 2, an extended semantic similarity measurement approach should count the resulting matches as partial match. Game and Sport both occupy the same thematic role (Theme) in the play function (full match), whereas Plasticine is defined as subrole of Essence in Figure 2 and therefore the match between Sport and Plasticine is only partial.

3.3 Thematic Roles as Feature Type

While the former section discused the relation between functional features and thematic roles, this section illuminates the question whether thematic roles can be regarded as feature type additionally to parts, functions and attributes as used in MDSM.

The question what can be done to or with something or what this thing affords to its environment seems to be a suitable way to model and categorize the world and especially artifacts [13][15][22][23]. As function is defined as the "role that an entity plays in serving the goal of an agent, or its role in the operation of a larger system such as a geology, ecology, or religion" [22, p. 2] and entities may have more than one function, it follows that an entity can play different roles within different contexts. Bob is the Agent of giving but the Recipient of receiving for example.

For entity classes this means that they can be subtypes of several thematic roles such as Agent and Recipient for Person (even at the same time: hurting oneself). This seams to contradict Sowa's idea of the placement of the thematic roles within formal ontology [2][6].

A possible solution is to regard thematic roles played by an entity class as directly connected to its functions. Person, in this sense it not a subclass of Agent and Recipient but an entity class with two functions (give and receive) that impose a certain role to the class. In other words Person is only an Agent in the situation of giving and not before or after a Theme is given to a Recipient. Sowa would argue that Person is no kind of Participant at all, but Giver and Receiver are (Giver is always an Agent for example). From this point of view thematic roles cannot be regarded as an additional kind of feature but have to be directly assigned to functions. One can even argue that in a valid model each subtype of a thematic role can only have one function namely the one that makes it an Agent for example.

Nevertheless another argumentation is possible that regards thematic roles as a way to describe the potential (potential ability) of class members. In this case thematic roles can be regarded as feature type that describes the tendency of how entities of a certain class behave. Stadium and Sport arena for example share the thematic role feature Location. Thus, all their members tend to behave as locations in associated functions, which make both entity classes seem to be similar. In cases where an entity class is described by more than one role feature the same conclusion is possible. Persons for example are entities that can behave either as Actors or Recipients but not as points in time or paths. Entity classes that have thematic role features in common can be therefore regarded as more similar than classes that differ in their role features.

3.4 Thematic Roles and Transformational Categories

As example for functions, Rodriguez and Egenhofer [1] argue that the class College has the functional feature educate. The function educate(x,y) can be interpreted either in a way that the college educates students or that the college is the location where students are educated. The corresponding conceptual graph for the former interpretation is:

[Building: Collage]←(Agnt)←[Education]→(Rcpt)→[Student:{*}]

The corresponding conceptual relation of Agent Agnt(Act, Animate) "restricts the usage to an active animate entity that voluntarily initiates an action" [2, p. 508). On the one hand this would mean that College is only the location where education takes place. On the other hand this is a classical example for metonymy [24] and reflects a human way of thinking and categorizing. Other examples for Metonymy in class definitions are the functions perform and present that are defined as features of Theater by Rodriguez and Egenhofer [1]. Again two interpretations are possible: a non-metonymic interpretation where Theater is the location where a group of actors present (or perform) a play and a metonymic interpretation where the Theater stands for the actors that present a play to the visitors. In the former case Theater can be regarded as subtype of Location (if we accept roles as feature types) and in the second case as of Agent. Theater specifies a change in the selection restriction of present(x, y) in a way that first Theater stands metonymical for a group of actors and than in a second step the group is regarded as a single Agent [15], which means that when Theater is defined as a transformational concept there is no contradiction in Agnt(Perform, Theater). By using thematic roles, an ontology engineer can restrict possible interpretations to the intended one.

3.5 Requirement for a Thematic Role Sensitive Similarity Measurement

An extended (feature-based) semantic similarity measurement theory that is able to deal with polysemy of functional features, metonymy within entity class names, potential behavior of class members (entities) and classes that are mostly defined by their role (role-governed) should support both views on thematic roles (as part of functional features and as feature type) and offer weightings for full and partial matches. Moreover, it should be able to integrate thematic roles in its context definitions.

4 Extending MDSM with Thematic Roles

In this section it is shown why and how an extended matching distance similarity measure (MDSM+TR) is able to fulfill the above requirements.

4.1 Extending the Entity Class Definition

MDSM requires a special class definition format, which can be regarded as a lightweight ontology [1]. To extend MDSM this class definition needs to be changed. As shown in Table 1, functions are defined as synstes, each synset containing

different words that represent the synonym symbols for a certain function. It is not possible to add the thematic roles as synset here, because they are not synonyms for functions. Instead, functions have to be referenced by pointers as this is done already for the entity classes in is_a for example. As minimum assumption for MDSM+TR, functions are defined by their name and the thematic role the possessing class (role_of_class) plays within this function. This definition is able to capture the relation between functions and thematic roles. To regard roles as an additional feature type another extension is necessary that defines the thematic role type as a list of roles as shown in Table 1.

Table 1. Differences between MDSM and MDSM+TR notation

BNF Notation: MDSM	**BNF Notation: MDSM+TR**

<entity_class>::= **entity_class** {
name: {syn_set}
description: <description>
is_a: <is-a>
part_of: <part_of>
whole_of: <whole_of>
parts: <parts>
functions: <functions>
attributes: <attributes>}
 	
<is_a>::= {}|{<pts_entity_classes>}
<part_of>::={}|{<pts_entity_classes>}
<whole_of>::={}|{<pts_entity_classes>}
<parts>::= {}|{<syn_sets>}
<functions>::= {}|{<syn_sets>}
<attributes>::= {}|{<syn_sets>}
<syn_sets>::={<syn_set>}|
<syn_sets>,{syn_set}
<syn_set>::= <word>|<syn_set>,<word>
<description>::= <word>|
<description><word>
<pts_entity_classes>::=<pointer>|
 	
<pts_entity_classes>,
<pointer>
	<entity_class>::= **entity_class** {
	…
	functions: <functions>
	thematic_roles: <functional_roles>
	…
	< functions>::= {}|{<pts_functions>}
	<pts_functions>}::=<pointer>|
	<pts_functions>,
	<pointer>
	<functional_roles>::={}|
	{<functional_role>}|
	<functional_roles>,
	{<functional_role>}
	<functional_role>::=<{x ∈ TR}>
	…
	
	<function>::= **function** {
	name: {syn_set}
	role_of_class: <functional_role>}
	…

4.2 Similarity Between Functional Features in MDSM+TR

In MDSM (Equation 2) $|C_1 \cap C_2|$ is defined by comparing the synset of each (functional) feature of c_1 to c_2. In other words, the (implicit) equal function used in MDSM examines function names (or sets of function names) and returns 1 for a match and 0 if the names do not match. For MDSM+TR this is not enough, because partial matches should be allowed too, and hence the thematic roles (role_of_class, see Table 1) have to be taken into account. Therefore the strength of a match has to be calculated (Equation 7) and the average of all matches has to be defined as weighting for $S_f(c_1,c_2)$.

$$\text{match}(tr_1, tr_2) = \frac{1}{1 + \text{arc\_distance}(tr_1, tr_2)} \tag{7}$$

The match function returns 1 for a full match (tr1=t2; e.g. match(Agent, Agent)) and $^1/_2$, $^1/_3$, $^1/_5$ or $^1/_7$ for the four different kinds of partial matches that are possible within the hierarchy of thematic roles (each is-a relation is regarded as one arc) [2]. In Equation 8 the weighting function ω_{ffp} is defined, that sums the strength of all matches within $C_1 \cap C_2$ and calculates the average ($c_1.ff_i.tr$ is the thematic role c_1 plays within the functional feature ff_i). The values for ω_{ffp} range between 1, if all matches are full matches and $^1/_7$ if only the function names match, but the roles are entirely different.

$$\omega_{\text{ffp}} = \frac{\sum_{i}^{|c_1 \cap c_2|} match(c_1.ff_i.tr, c_2.ff_i.tr)}{|\, c_1 \cap c_2 \,|} \tag{8}$$

Some cells of the thematic roles matrix contain two roles, in this case arc_distance is defined as 2 (e.g. Agent-Initiator-Effector).Equation 9 shows the MDSM+TR version of $S_f(c_1,c_2)$ whereas S_p and S_a remain as there are.

$$S_f(c_1, c_2) = \omega_{\text{ffp}} \cdot \frac{|\, C1 \cap C2 \,|}{|C_1 \cap C_2| + \alpha \cdot (c_1, c_2) \cdot |C_1 \setminus C_2| + (1 - \alpha \cdot (c_1, c_2)) \cdot |C_2 \setminus C_1|} \tag{9}$$

4.3 Similarity Between Thematic Role Features in MDSM+TR

In order to take thematic roles as additional feature type into account it is necessary to extend the overall similarity measure $S(c_1,c_2)$ by a weighting ω_{tr} and the similarity measurement for roles $S_{tr}(c_1,c_2)$ as described in Equation 10. The similarity function $S_t(c_1,c_2)$ is the same as in MDSM and each role can appear only one time per entity class.

$$S(c_1, c_2) = \omega_p \cdot S_p(c_1, c_2) + \omega_f \cdot S_f(c_1, c_2) + \omega_a \cdot S_a(c_1, c_2) + \omega_{t_r} \cdot S_{t_r}(c_1, c_2) \tag{10}$$

4.4 Thematic Roles and Context in MDSM+TR

In MDSM the weighting function ω_t is defined by variability or commonality and then normalized, so that the sum of the weightings is always 1. For P_f^v and P_f^c one has to decide whether the number of occurrences (o_i) of a certain functional feature within the domain of application is determined by its name or the combination of name and role. Partial matches can not be taken into account here, because this would violate the model of variability and commonality within MDSM. The author prefers the latter method because it reduces the effect of polysemous function names, increases variability (decreases commonality) and therefore strengthen the importance of functions within overall similarity. This is especially important for entity classes that are mostly defined by their functions (role-governed) and artifact classes (such as buildings or devices) in general.

$$\omega_{tr} = \frac{P_{tr}^{\ v}}{P_p^{\ v} + P_f^{\ v} + P_a^{\ v} + P_{tr}^{\ v}} \tag{11}$$

In the case where thematic roles are regarded as additional feature type, P_t^v and P_t^c do not need to be changed, but the weighting functions (6a, 6b, 6c and 8a, 8b, 8c in [1]) have to be extended by P_{tr}^v or P_{tr}^c as this is demonstrated for variability in ω_{tr} (Equation 11).

5 Theater, Sport Arena and Guitar

This section presents some measurement examples from a test-ontology and discusses the different results between MDSM and MDSM+TR.

5.1 Experiment

To prove the idea of the thematic roles extension semantic similarity between the entity classes Theater, Sport arena (both taken from Table 2 of [1]) and Guitar is measured using MDSM and MDSM+TR. Theater is defined in to ways: one that regards Theater as Actor of the functional features perform and present and another where Theater plays the role of a Location (see Table 2).

Table 2. Feature description for Theater, Sport arena and Guitar

Entity Class	Parts	Functions	Attributes	Roles
Theater_1	Dressing room Entrance hall Foundation Orchestra Roof Spectator stands Stage Ticket office Wall	Perform(L) Present(L) Recreate(L)	Architectural properties Ext. material construction Height Location Name Owner type Structure type User type	Location
Theater_2	As above	Perform(A) Present(A) Recreate(L)	As above	Agent Location
Sport arena	Court Dressing room Foundation Roof Spectator stands Wall	Play(L) Practice(L) Recreate(L)	Architectural properties Ext. material construction Height Location Name Owner type Structure type User type	Location
Guitar	Body Strings	Play(I) Practice(I) Recreate(I)	Type Material Color	Instrument

The context is defined as C=<{recreate, {}> which means that the domain of application contains the four entity classes displayed in Table 2. It has to be emphasized that Theater_1 and Theater_2 are both taken into account for the calculation of weightings which decreases variability within the domain. Moreover Guitar (which is used here as a kind of false-positive for the similarity calculation within functional features in MDSM and therefore contains the same functions as Sport arena) is specified by few features only which additionally decrease variability.

The aim of this similarity measurement experiment is to show how MDSM+TR behaves in certain situations in comparison to MDSM. Theater_1 and Theater_2 will never be part of the same ontology and same context in real world measurements for example.

Table 3. Some relevant values from the similarity measurement with MDSM and MDSM+TR

Model	c_1 versus c_2	P_f^v	P_{tr}^v	$S_f(c_1,c_2)$	$S_{tr}(c_1,c_2)$	$S(c_1,c_2)$
MDSM	Theater_1 vs. Theater_2	0.4	--	1.0	--	**1.0**
MDSM+TR	Theater_1 vs. Theater_2	0.7	0.58	0.71	0.66	**0.82**
MDSM	Theater_1 vs. Sport arena	0.4	--	0.33	--	**0.66**
MDSM+TR	Theater_1 vs. Sport arena	0.7	0.58	0.33	1.0	**0.7**
MDSM	Theater_2 vs. Sport arena	0.4	--	0.33	--	**0.66**
MDSM+TR	Theater_2 vs. Sport arena	0.7	0.58	0.33	0.66	**0.61**
MDSM	Guitar vs. Sport arena	0.4	--	1.0	--	**0.32**
MDSM+TR	Guitar vs. Sport arena	0.7	0.58	0.43	0.0	**0.14**

5.2 Discussion of the Results

The results presented in Table 3 show some relevant results from the similarity measurement using MDSM and MDSM+TR, where S is the overall similarity, S_f and S_{tr} are the similarities for the functional features and thematic roles and P_t^v and P_{tr}^v are the results for variability of functional features and thematic roles. The functional feature extension of MDSM+TR tends to decrease similarity because it introduces more information about functions. If name and role_of_class are equal for the compared functional features the results between MDSM and MDSM+TR do not differ (Theater vs. Sport arena), but are decreased the more different the roles of the entity classes within the compared functions are. Therefore S_f(Theater_1, Theater_2) is not 1.0 but 0.71 in the MDSM+TR approach and 0.43 instead of 1.0 for S_f(Guitar, Sport arena). The functional features of Guitar and Sport arena have nothing more than their names in common (polysemous function names).

The thematic role feature type offers an additional possibility to compare entity classes and is therefore able to increase or decrease similarity. On the one hand in S(Theater_2, Sport arena) the overall similarity is decreased because Theater_2 does not only play the role of a Location but can be regarded as an Agent too. On the other hand S(Theater_1, Sport arena) is increased by S_{tr}(Theater_1, Sport arena) because the compared classes both play the role of a Location.

In border cases such as S(Guitar, Sport arena) the differences between MDSM and MDSM+TR may be very heigh (MDSM: 0.32; MDSM+TR: 0.14) but in general the results should not vary more than between 5-20%. The thematic role feature type

similarity $S_{tr}(c_1,c_2)$ has more impact on the model than the role-based partial matches for $S_f(c_1,c_2)$. Therefore the latter one can be regarded more as a refinement than an extension to the MDSM theory.

6 Conclusions and Future Work

Thematic roles can be easily integrated into MDSM and improve the theory to fulfill the requirements defined in this paper. The resulting MDSM+TR is able to handle polysemous functional feature names and metonymy within entity class names. By taking thematic roles as an additional feature type into account, MDSM+TR is able to measure similarity based on the idea that entity classes whose members behave in a common way (play a certain role) are similar. Thematic roles are more than just another feature type such as parts, functions and attributes, because they come with a very generic theory of participation that adds more structure to the entity class description (and the functional features). While the names (symbols) and the meaning of other features may differ from ontology to ontology, thematic roles are fixed within Sowa's formal ontology and therefore are able to restrict possible interpretations. The ontology design process has fundamental influence on the similarity measurement and as argued in Goldstone and Son [11] all entity classes can be made similar to each other by adding features such as lessthan5000pound or colored for example. Moreover we do not measure similarity between concepts (in our mind) or real world entities but between representations (models); what sounds trivial first, is a fundamental restriction to all assumptions made by using computational theories of similarity. Even within a single ontology granularity can vary between the concept specifications, which directly influence the resulting similarity. All we can state from this kind of measurement is that according to the examined ontology c1 and c2 are similar to a certain degree represented by a numerical value. It is up to the user to decide what similarity value is sufficient for a certain task. MDSM uses a lightweight ontology that primarily consists of meaningless labels without any relation to each other or axioms, which additionally increases the influence of the ontology engineer and makes the measurement very design and granularity dependant. Nevertheless similarity is an important theory for information retrieval and discovery within ontologies, because it is not only able to return classes suitable for a certain task but offers also a ranking. The extension presented in this paper is a first step to a more semantic comparison of distinguishing features (functional and thematic role features) as proposed by Rodriguez and Egenhofer.

A lot of work remains to be done such as human subject testing. Moreover the theory presented here only takes the participant hierarchy into account to express partial matches leaving the verb categories beside. Future work is necessary to analyze how this aspect can be added to the model. The six verb categories are not a final set and on a very abstract level, Sowa [6] argued that they can be divided into more categories if necessary. For the geo domain it would be of special interest to analyze the temporal and spatial categories and create additional sub roles if necessary.

References

1. Rodríguez, A. and Egenhofer M. J. Comparing Geospatial Entity Classes: An Asymmetric and Context-Dependent Similarity Measure. *International Journal of Geographical Information Science*, 18(3): 229-256, 2004.
2. Sowa, J. F. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing Co., Pacific Grove, CA., 2000.
3. Moravcsik, J. M. What Makes Reality Intelligible? Reflections on Aristotle's Theory of Aitia. *Aristotle's Physics: A Collection of Essays*. Ed. Lindsay Judson. New York: Clarendon, pp. 31-48, 1991.
4. Dick, J. A conceptual, case-relation representation of text for intelligent retrieval, PhD thesis, Department of Computer Science, University of Toronto. Published as technical report CSRI-265, 1991.
5. Pustejovsky, J. *The Generative Lexicon*. Cambridge/London: MIT Press, 1995.
6. Sowa, J. F. Processes and Participants. In Peter Eklund, Gerard Ellis, and Graham Mann, editors, *Conceptual Structures: Knowledge Representation as Interlingua, number* 1115 in Lecture Notes in Artificial Intelligence. Springer-Verlag, 1996.
7. Somers, H. L. *Valency and case in computational linguistics.* Edinburgh: Edinburgh University Press, 1987.
8. Masolo, C., Vieu, L., Bottazzi, E., Catenacci, C., Ferrario, R., Gangemi, A., Guarino, N. Social Roles and their Descriptions. In: *Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning*, 2004.
9. Tversky, A. Features of similarity. *Psychological Review*, 84(4): 327-352, 1977
10. Goldstone, R. L. The role of similarity in categorization: providing a groundwork. *Cognition*, 52: 125-157, 1994.
11. Goldstone, R. L. and Son J. Y. Similarity. Cambridge *Handbook of Thinking and Reasoning*. K. Holyoak and R. Morrison. Cambridge, Cambridge University Press, 2004.
12. Gärdenfors, P. *Conceptual Spaces - The Geometry of Thought*. Cambridge, MA, Bradford Books, MIT Press, 2000.
13. Gibson, J. *The Ecological Approach to Visual Perception*. Boston, Houghton Mifflin Company, 1979.
14. Goldstone, R. L., Medin, D. L. and Halberstadt, J. Similarity in context. *Memory and Cognition*, 25: 237-255, 1997.
15. Markman, A. B. and Stilwell, C. H. Role-governed categories. *Journal of Experimental and Theoretical Artificial Intelligence*, 13: 329-358, 2001.
16. Gentner, D., and Kurtz, K. Learning and using relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman & P. W. Wolff (Eds.), *Categorization inside and outside the lab*. Washington, DC: APA (in press)
17. Wittgenstein, L. *Philosophical investigations, trans. G.E.M. Anscombe*. New York: MacMillan, 1968.
18. Jones, D. M. and Love B. C. Beyond common features: The role of roles in determining similarity. *CogSci 2004 - 26th Annual Meeting of the Cognitive Science Society*, Chicago, US, 2004.
19. Barsalou, L. W. Ad hoc categories. Memory & Cognition 11(3) (1983) 211-227
20. Kuhn, W. Modeling the Semantics of Geographic Categories through Conceptual Integration. GIScience 2002, Boulder, CO, USA, D. Mark, Editor. Springer: Berlin, pp. 108-118, 2002.
21. Markman, A. and Gentner D. Structural Alignment during Similarity Comparisons. *Cognitive Psychology*, 25: 431-467, 1993.

22. Barsalou, L.W., Sloman, S.A, and Chaigneau, S.E. The HIPE theory of function. In L. Carlson & E. van der Zee (Eds.), *Representing functional features for language and space: Insights from perception, categorization and development*. Oxford: Oxford University Press, (in press).
23. Khoshafian, S. and Abnous, R. *Object Orientation: Concepts, Languages, Databases, and User Interfaces*. New York, John Wiley & Sons, 1990.
24. Radden, g. and Kövecses, Z. Towards a Theory of Metonymy. In *Panther*, Klaus-Uwe; Radden, Günter (eds.), 1999.