Measurement of Situation Similarity

Chuong C. Vo, Torab Torabi, and Seng W. Loke
Department of Computer Science & Computer Engineering
La Trobe University, VIC 3086, Australia
Email addresses: {c.vo, t.torabi, s.loke}@latrobe.edu.au

Abstract

Context-aware computing applications often perform an assessment of similarity between the user's current situation and the predefined or learned situations in order to trigger appropriate adaptation rules. Because future situations are often not exactly matching with past ones (e.g., time continuously changes), the assessment of similarity between situations is not a straightforward task. This paper reviews approaches to this problem and investigates various measures for similarity in the context of situation similarity. As a result, we propose a novel measurement of situation similarity. The main different between our approach from the existing ones is that we take into consideration of context in which the comparison of two situations occurs. The evaluation shows that our method achieves good results in comparison with others.

Keywords: Situation Similarity, Context Matching

1 Introduction

...

2 Concepts

2.1 Context

We adopt an operational definition of *context* by Dey [1]: "*Context is any information that can be used to characterise the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves*". The typical *context attributes* are location, time, environmental conditions, user profile, activity, nearby objects, devices, and nearby people. It is obvious that which context attributes being chosen to used depend on individual applications.

Each context attribute can be decomposed into different levels of generalisation. For example, the time context as shown in Figure 1 can be decomposed into hour, time-of-day, date, day-of-week, time-of-week, month, season, and year.

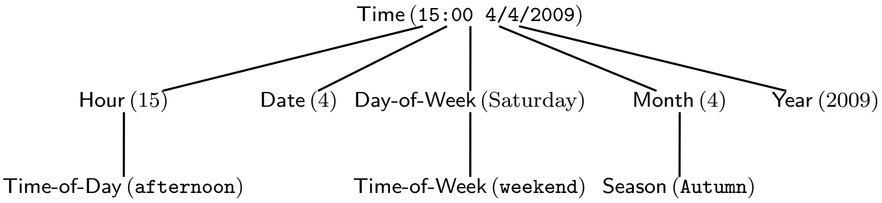

Figure 1: The decomposition of the time context

2.2 Situation

Situation of a user is characterised by context attributes. In a specific situation-sensitive application, situations are often pre-known, manually pre-defined, labelled by name phrases such as `Meeting` and `Sleeping`. Situations are sometimes called *high-level context abstractions* [2]. There can be many representations for a situation. For example, Loke [3] identified six different representations (perhaps more than six) of a the Meeting situation:

- You are with someone, and have an entry for a meeting in your diary;
- You are with your colleagues with filled coffee cups;
- You are in the room having the weight of the floor above a threshold.
- You are in the room with a projector switched on, the room lights on, and an Power Point application running on a PC.
- You are in the room with a noise level above a threshold.
- You are in the room with the presence of many people (detected by cameras in the room).

Early approaches such as [4, 5] use formal logics to describe and represent situations. Even though formal logics based presentations provide formality for logically specifying the situations, they are error-prone because of the incompleteness and ambiguity of contextual information [2]. The limited reasoning performance of these approaches further reduces their usability in real world mobile applications.

3 Concepts of Similarity Measurement

3.1 Notion of Similarity Measurement

The theory of similarity has its roots in philosophy and psychology and was established to determine why and how entities are grouped into categories, and why some categories are comparable to each other while others are not. As argued by Goodman [6], there is no global and application independent law on how similarity is measured. There is even no single definition of what similarity measures [7].

There are two main types of similarity: *dimensional similarity* (or *properties-based similarity*) and *global similarity* (or *taxonomy-based similarity*). For global similarity, the objects compared are richly structured, or, in other words, each object is hierarchically constituted of sub-objects linked by heterogeneous relations.

According to [8], a framework of similarity measurement should conform the following advantages:

Consistency be coincident with the existing experimental facts, common senses, and practical experiences;

Conciseness should have few (even no) prior conditions or hypothesis;

Compatibility be compatible with all the reasonable ingredients of the existing similarity models;

Simplicity be as simple as possible;

Universality should achieve sufficient universality.

The three intuitions about similarity measurement are proposed in [9]:

Intuition 1 The similarity between A and B is related to their commonality. The more commonality they share, the more similar they are;

Intuition 2 The similarity between A and B is related to the differences between them. The more differences they have, the less similar they are;

Intuition 3 The maximum similarity between A and B is reached when A and B are identical, no matter how much commonality they share.

3.2 Notion of Situation Similarity

As discussed above, because situations are represented by context attributes, the evaluation of similarity between situations is reduced to the evaluation of similarity between context attributes. Further, because context attributes are structural data objects, measuring similarity between context attributes now becomes measuring similarity between data objects.

Requirements of a situation similarity measure:

- can account for complex semantics: the similarity measure has to model properties as well as relations and should provide a structured similarity measure; and
- must support similarity at a conceptual level, not single instances of features;

4 Basic Models of Similarity

4.1 Geometric Models

Geometric models are based on the notion of multi-dimensional vector spaces. Dimensions are used to describe features of objects and concepts. Each dimension is a set of ordered values. This enables the comparison of two features of the same quality, e.g. cold is more similar to cool than to hot.

Because each object is represented by a point in a multidimensional feature space, similarity is then inversely related to the distance between points in the space. The most commonly used similarity measures are the Minkowski distance measures (Equation 1):

$$d(O, O') = \left(\sum_{i=1}^{n} |o_i - o'_i|^r \right)^{1/r}, \tag{1}$$

where n is the number of dimensions, o_i is the value of dimension i for object O and o'_i is the value of dimension i for object O'. $r = 1$ results in the City-block distance and $r = 2$ in the Euclidian distance. Similarity is a linear decaying function of $d(O, O')$.

The geometric similarity model exposes the three metric axioms:

- *Minimality*: $D(A, B) \geq D(A, A) = 0$;
- *Symmetry*: $D(A, B) = D(B, A)$; and
- *Triangle inequality*: $D(A, B) + D(B, C) \geq D(A, C)$. It equals to $S(A, B) + S(B, C) \leq S(A, C)$.

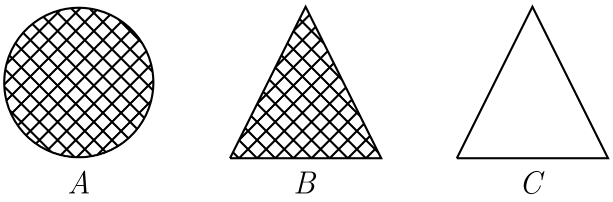

Figure 2: Counterexample of triangle inequality

Unfortunately, violations of all three assumptions are empirically observed [10]. For example, in Figure 2, A and B are similar in their shades, B and C are similar in their shapes, but A and C are absolutely not similar. Another famous example for similarity judgments that violate the triangle inequality: A lamp is similar to the moon (providing light) and the moon is similar to a football ball (in their shapes); but a lamp is not similar to a football ball.

4.2 Feature Models

In feature models, objects or concepts are represented via an unstructured set of features. For example, a Meeting happened in a room may be described by the features: crowed, silent, and working hours. Or a person sleeping in a room may be described by the feature: silent, dark, and no movement.

Similarity is determined by common and distinctive features of the objects compared [10, 11]. The contrast model is:

$$S(A, B) = \alpha |A \cap B| - \beta |A - B| - \gamma |B - A|,$$

and the ratio model is:

$$S(A, B) = \frac{|A \cap B|}{\alpha |A \cap B| + \beta |A - B| + \gamma |B - A|},$$

where $|A \cap B|$ is the cardinality of the intersection of set A and set B; $|A - B|$ is the the cardinality of the subtractions of set A and set B; α, β, and γ are weight factors. Because these factors are not equivalent, the feature models are asymmetric.

Feature models assess only entire feature matches and cannot detect partial similarity between features, i.e. features such as *institutional place* and *meeting place* do not match and are as dissimilar as *institutional place* and *teaching place.* In contrast to the geometric model, feature models cannot detect intra-dimensional feature similarity.

4.3 Network Models

Network models are graph-based and use semantic networks for knowledge representation. Semantic networks are composed of labelled nodes and edges. Nodes represent units of knowledge (e.g., objects, concepts or properties). Edges link nodes with each other and represent the relations between them explicitly. Similarity in network models is computed based on the shortest path in a network (e.g., [12, 13]). Rada et al. [13] proposed DISTANCE to compute similarity based on the shortest path between concepts in a semantic network with taxonomic relations. Rodriguez [14], Rodriguez and Egenhofer [15] proposed a semantic similarity measure based on the feature similarity of a concept's semantic neighbourhood. A semantic neighbourhood of a concept c is defined as the set of concepts whose network distances to c are equal to or less than r, which is the radius of the neighbourhood. Two concepts are more similar if their semantic neighbourhoods are more similar. SimRank [16] is a similarity measure which computes similarity of two objects based on their relationships with other objects. That is, two objects are more similar if they are related to more similar objects. The strength of the network approach is the representation of relations between concepts. Problem is over-reliance on structural details of the network. The problem with this approach is that it relies on the notion that edges in a taxonomy represent uniform distances (i.e., it assumes that all semantic links are of equal weight).

4.4 Alignment Models

While geometric and feature models search only for matching elements, alignment models [17] also account for whether these matching elements align or not. Alignable matches increase the similarity more than non-alignable matches.

4.5 Transformational Models

The similarity of two objects is assumed to be inversely proportional to the number of operations required to transform one object so as to be identical to the other [18]. The main problem is to find suitable sets of transformations. The transformational model holds minimality, triangle inequality axioms, but is asymmetric.

4.6 Logic Models

Similarity is measured by determining the overlap of a query concept definition and concept definitions in description logics [19]. It requires concepts to be described by description logics.

4.7 Hybrid Models

The hybrid similarity measure such as [20] combines the geometric structure of conceptual spaces with the relational structure of semantic networks. Each concept is described by its properties within a conceptual space. Then, concepts are related via links in a network structure to other concepts which are again described in conceptual spaces.

5 Approaches on Situation Similarity Measures

5.1 Ontology-Based Similarity

Situations are represented by nodes in a *situation ontology* [21, 22]. The similarity measure of concepts in an ontology is based on their distances within the ontology. Give a situation ontology. Obviously, Public Meeting is more similar to Private Meeting than to Sleeping. The ontology distance could be defined as the shortest path going through a *common ancestor*. The pseudo-code algorithm is as follows [23]:

- gen_A = all transitive generalisations of the situation A;
- gen_B = all transitive generalisations of the situation B;
- from $gen_A \cap gen_B$, determine the most recent common ancestor (MRCA);
- ontology distance $d(A, B)$ = count the length of the path from A to B via MRCA.

The problem of the ontology distance is that it is highly dependent on the construction of the ontology.

5.2 Information-Theoretic Approaches

To address the problem of the ontology distance based approaches, researchers have proposed information-theoretic entropy measures (e.g., [9, 24]). Specifically, Resnik [24, 25] argues that a node (e.g., concept, object, word) is defined by its members. When using an explicit ontology like WordNet, the set of members is equivalent to the descendants (hyponyms) of a node. The information of a node is defined as the probability of finding the particular set of descendants, its entropy as the negative log of that probability. The similarity is now defined as

$$sim(A, B) = \frac{2 \log P(MRCA(A, B))}{\log P(A) + \log P(B)},$$

where MRCA is the most recent common ancestor of nodes A and B. Intuitively, this measure specifies similarity as the probabilistic degree of overlap of descendants between two objects. The following algorithm:

- U = the total number of objects;
- Find the most recent common ancestor (MRCA) of A and B;
- P(A) = (number of specialisations of A) / U;
- P(B) = (number of specialisations of B) / U;
- P(MRCA(A, B)) = (number of specialisations of MRCA) / U;
- $sim(A, B) = (2 \log P(MRCA(A, B)))/(\log P(A) + \log P(B))$.

example...

5.3 Vector Space Approaches

Each object is represented as a vector of features in a k-dimensional space and compute the similarity by measures such as cosine or Euclidean distance. Here k is the number of unique object attributes or relations of the object. The similarity between two objects' vectors is now simply defined as their inner product. The pseudo-code algorithm is:

- Determine vector x from the object parts of A;
- Determine vector y from the object parts of B;
- $sim(A, B) = |xy|/(|x| * |y|)$.

As an example, consider the object *chair*, which has four *legs* and one *back* to which it has a *has-part* relation as well as a *room office* to which it has a *is-in* relation. The chair vector [4, 1, 1] would represent the chair in the space with the dimensions [*has-part legs*, *has-part back*, *is-in office*]. Clearly, this type of "vectorisation" is problematic as it, for example, does not capture that the dimensions *has-part legs* and *has-part back* are (semantically) closer related to each other than to *is-in office.* However, it has the advantage of being computationally cheap.

5.4 Edit Distance - Transformational Model

The similarity between strings is often described as the edit distance. It is the number of changes necessary to turn one string into another. Here a change is typically defined as either the insertion of a symbol, the removal of a symbol, or the replacement of one symbol with another. In our case, therefore, we calculate the number of transformation steps needed to turn one object into another object. In other words, we count the number of insert, remove, and replacement operations of attributes, attribute values, relationships, or relationship types. In a first version we assume equal costs (=1) for each of the transformations. In an alternative implementation, we weigh each transformation type with a value that represents the "real" costs. For example, is the replacement transformation comparable with a deletion procedure followed by an insertion procedure? Hence, we could argue that the cost function c would have the following behaviour:

$$c(deleting) + c(inserting) >= c(replacing).$$

Using this assumption we calculate the worst (i.e., most costly) case for a transformation from A to B by replacing all object parts of A with object parts of B, then deleting the rest of the object parts of A, and inserting additional object parts of B into A. The worst case cost is then used to normalise the edit distance to a similarity. The overall algorithm looks as follows:

- Determine parts (attributes/relationships) of A;
- Determine parts of B;
- Compute number of actual transformation steps (replace, insert, delete) from A to B;
- Compute worst case cost for the procedure (as mentioned above);
- Relative edit distance = (number of transformation steps) / (worst case costs).

5.5 Full-Text Retrieval Method

Probably the most often-used similarity measure comes from the information retrieval literature and compares two documents by using a weighted histogram of the words they contain. Specifically, the measure works as follows: it counts the frequency of occurrence of each term in a document in relation to the term's occurrence frequency in a whole corpus of documents. The resulting word counts are then used to compose a weighted term vector describing the document. The similarity between two documents is now computed as the cosine between their respective weighted term vectors.

In our case, we create a (text) document for each object like a object description. Then we compute their similarity.

In the following, we will examine spatial similarity as a representative. Other context attributes which share the same structure can be applied by the same approach.

"Geographic features are distinguished via their geometric (the geometric structure of conceptual spaces) and thematic data."

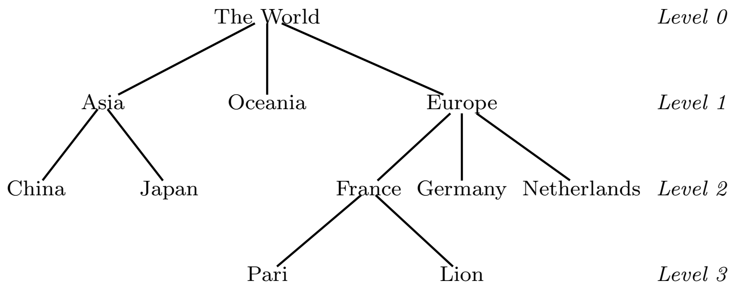

Figure 3: A simplified geographical hierarchy

Makkonen et al. [26] introduce a measurement for spatial similarity, they use a geographical hierarchy to make connections between spatial entities. Figure 3 shows a simplified geographical hierarchy containing a number of nodes. Each node on the tree stands for a real-life place. To measure the geographical similarity of two places l_1 and

l_2, they compare the length of the common path to the two nodes (representing the places) to the sum of the lengths of the paths to these nodes:

$$\mu_g(l_1, l_2) = \frac{path(l_1 \cap l_2)}{path(l_1) + path(l_2)}. \tag{2}$$

The following similarities are computed by using (2).

$$\begin{aligned}
\mu_g(Paris, Lion) &= \frac{path(Paris \cap Lion)}{path(Paris) + path(Lion)} = \frac{2}{3+3} \approx 0.33. \\
\mu_g(Paris, France) &= \frac{path(Paris \cap France)}{path(Paris) + path(France)} = \frac{2}{3+2} = 0.4. \\
\mu_g(France, Paris) &= \frac{path(France \cap Paris)}{path(France) + path(Paris)} = \frac{2}{2+3} = 0.4.
\end{aligned}$$

In [27], a context matching algorithm is introduced. Their approach for measuring similarity between locations is almost similar to [26]. The only difference is that they conventionally define $\mu_g(l_1, l_2)$ equal to 1 if l_1 covers l_2. For example, $\mu_g(Paris, France) = 0.4$ but $\mu_g(France, Paris) = 1$. This should be understood that if the location is *Paris*, then it is absolutely in *France*; and if the location is *France*, then it is unsure to be *Paris*.

The formula (2) has three limitations. Firstly, (2) is not applicable for evaluating the similarity between two identical places; (2) gives $\mu_g(l, l) = 0.5$ but practically $\mu_g(l, l) = 1$. Secondly, (2) never returns a similarity of two places being greater than 0.5 because it redundantly attaches the length of common path in the sum $path(l_1) + path(l_2)$. Finally, from the examples above, we see $\mu_g(Paris, Lion) \neq \mu_g(Paris, France)$. This is not practical because *Lion* is in *France*, hence logically $\mu_g(Paris, Lion) \geq \mu_g(Paris, France)$.

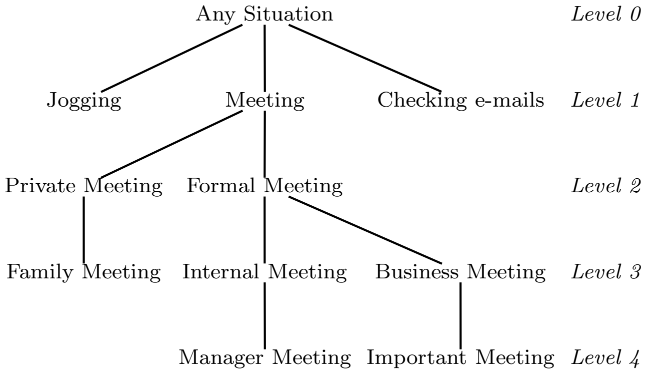

Figure 4: A Taxonomy of Situational Context

Anagnostopoulos et al. [28] propose a concept similarity measure. Specifically, consider a concept $D \in \Theta$ of a given taxonomy H, let also H_D be the set of superconcepts of D (plus itself), thus $H_D = \{D_j | (D \subseteq D_j) \cap \{D\}\}$. For instance, for $D = BusinessMeeting$, then $H_D = \{BusinessMeeting, FormalMeeting, Meeting, AnySituation\}$ (See Figure 4). The similarity measure of two concepts D and C belong to the same taxonomy H is defined as:

$$sim(D, C) = \frac{|H_C \cap H_D|}{|H_C \cap H_D| + \alpha |H_D \setminus H_C| + \beta |H_C \setminus H_D|}, \tag{3}$$

where $|H|$ is the cardinality of the H set. The factors α, β refer to the weights for common and different features respectively and $\alpha+\beta = 1$. Let consider some examples with $\alpha = 0.5, \beta = 0.5$:

$$\begin{aligned}
sim(Paris, Lion) &= \frac{|H_{Paris} \cap H_{Lion}|}{|H_{Paris} \cap H_{Lion}| + 0.5|H_{Paris} \setminus H_{Lion}| + 0.5|H_{Lion} \setminus H_{Paris}|} = \\
&= \frac{3}{3 + 0.5 * 1 + 0.5 * 1} = 0.75. \\
sim(Paris, France) &= \frac{|H_{Paris} \cap H_{France}|}{|H_{Paris} \cap H_{France}| + 0.5|H_{Paris} \setminus H_{France}| + 0.5|H_{France} \setminus H_{Paris}|} = \\
&= \frac{3}{3 + 0.5 * 1 + 0.5 * 0} \approx 0.86. \\
sim(France, Paris) &= \frac{|H_{France} \cap H_{Paris}|}{|H_{France} \cap H_{Paris}| + 0.5|H_{France} \setminus H_{Paris}| + 0.5|H_{Paris} \setminus H_{France}|} = \\
&= \frac{3}{3 + 0.5 * 0 + 0.5 * 1} \approx 0.86.
\end{aligned}$$

Qin et al. [29] represent context object as an n-dimension vector: $C = (c_1, c_2, \ldots, c_n)$, where $c_i, (i \in \{1, \ldots, n\})$ is a context type (e.g., Activity, Location) ranging from -1 to 1. The similarity degree between two context objects C_1 and C_2 is calculated by using the *Pearson's correlation coefficient* as:

$$Sim(C_1, C_2) = \frac{C_1 \cdot C_2}{||C_1|| \times ||C_2||} = \frac{\sum_{i=1}^{n}(\alpha_i \beta_i)}{\sqrt{\sum_{i=1}^{n}(\alpha_i^2)}\sqrt{\sum_{i=1}^{n}(\beta_i^2)}},$$

where $C_1 = (\alpha_i), C_2 = (\beta_i), i \in \{1, ldots, n\}$.

Loke [30] defines a *situation S* as a *rule*. For example, the rule representing a situation of *in-meeting(E)* is "**If** *in-meeting(E)* **Then** *with-someone(E)* AND *has-entry-for-meeting-in-diary(E)*". While the rule representing the situation of *with-someone(E)* is "**If** *with-someone(E)* **Then** *location*(E, L), people-in-room*(L, N), N* > 1", and the rule representing the situation of *has-entry-for-meeting-in-diary(E)* is "**If** *has-entry-for-meeting-in-diary(E)* **Then** *current-time*(T1), diary*(E, 'Meeting', entry(StartTime, Duration)), within-interval(T1, StartTime, Duration)*".

Then, if the current contextual information about an entity E holds in S, then E is in the real world situation represented by S.

6 Corrected Measurement of Context Similarity

We propose a corrected method that measures the geographical similarity of place l_1 over place l_2 by comparing their common path and the path to l_1:

$$\mu_g(l_1, l_2) = \frac{path(l_1 \cap l_2)}{path(l_1)}, \tag{4}$$

where $path(l_1 \cap l_2)$ represents the length of the common path of l_1 and l_2. Obviously, (4) gives $\mu_g(l_1, l_1) = 1$. Now, let us recompute the previous similarities using (4).

$$\mu_g(Paris, Lion) = \frac{path(Paris \cap Lion)}{path(Paris)} = \frac{2}{3} \approx 0.66.$$

$$\mu_g(Paris, France) = \frac{path(Paris \cap France)}{path(Paris)} = \frac{2}{3} = 0.66.$$
$$\mu_g(France, Paris) = \frac{path(France \cap Paris)}{path(France)} = \frac{2}{2} = 1.$$

Our proposed measurement addresses three limitations of the previous method aforementioned via only a unified formula. It is applicable for every cases.

6.1 Extending the geographical hierarchy

We extend the geographical hierarchy to include more granular levels of locations such as landmark, building, and room as depicted in Figure 5.

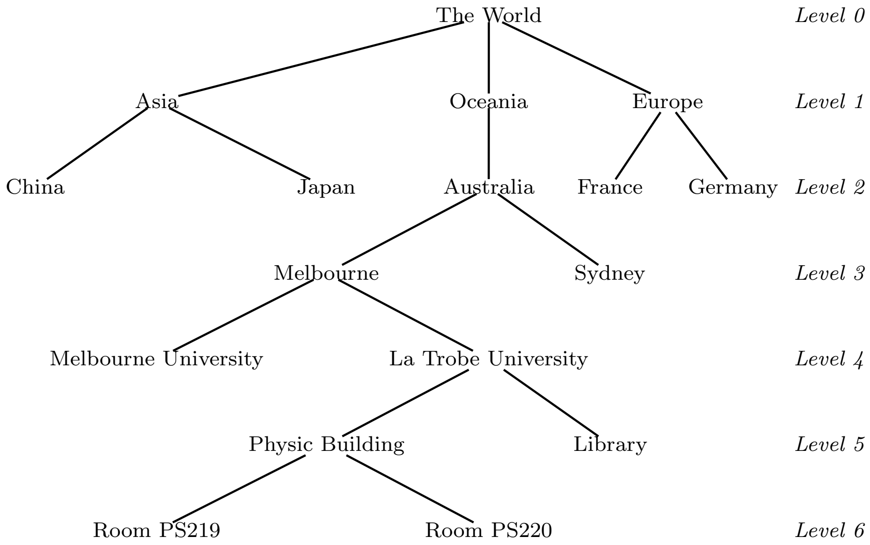

Figure 5: A more granular geographical hierarchy

Now, by applying (4), we can measure the similarity of two places *PS219* and *PS220*:

$$\mu_g(PS219, PS220) = \frac{path(PS219 \cap PS220)}{path(PS219)} = \frac{5}{6} \approx 0.83.$$

6.2 Integrating Other Aspects Into Similarity Measurement

In some cases, while the *geographical similarity* between two places is insignificant, they may be identical with regards to place types. For example, two seminar rooms, one is at a Paris University in France and another is at La Trobe University in Australia, their geographical similarity would equal zero whereas their similarity in term of place types should equal to 1 because they are all rooms mainly for seminar activities. Therefore, we propose that the final similarity between two places are the sum of their geographical and place-type based similarities respecting to some weights assigned to each similarity.

$$\mu(l_1, l_2) = w_g * \mu_g(l_1, l_2) + w_t * \mu_t(l_1, l_2), \tag{5}$$

where w_g and w_t are weights mentioned above (note that, $w_g + w_t = 1$), $\mu_t(l_1, l_2)$ is the place type based similarity between l_1 and l_2.

"Weight of objects: Different objects have different effects on the systematic similarity degree, as is quantified by weight, which represents the importance of an object with regard to its contribution to the systematic similarity. For example, in an exploration of the resemblance of human faces, the eyes may be assumed to be of higher importance than other facial organs. The weight of an entity can be considered as a function of its attributes. One may think that weight is constant for a given entity; however, in many scientific fields, weight is vulnerable to change with context, perspective, and preference, which causes the mutability of a similarity degree judgment. In practice, weight is usually quantified by some combination of statistical measure"

A place can belong to numbers of types. The more types a place belongs to, the more specific it is. We refer place types as properties of a place. The set of place types would be "*private place*", "*public place*", "*indoor place*", "*outdoor place*", "*institutional place*", "*shopping place*", "*meeting place*", "*teaching place*", so on. To measure the place type-based similarity of place l_1 over place l_2, we employ the *feature contrast model* [11], wherein similarity is determined by common and distinctive place types of the places compared. The similarity of compared places is assumed to increase with its common place types and decrease with its different place types. Particularly,

$$\mu_t(l_1, l_2) = \frac{card(l_1 \cap l_2)}{card(l_1)}, \tag{6}$$

where $card(l_1 \cap l_2)$ is the cardinality of place types that l_1 and l_2 have in common, $card(l_1)$ is the cardinality of place types that l_1 belongs to. For example, the place types that the room *PS219* belongs to is {*public*, *indoor*, *institutional*, *meeting place*} and that the room *PS220* belongs to is {*public*, *indoor*, *institutional*, *teaching place*}. We have:

$$\mu_t(PS219, PS220) = \frac{card(PS219 \cap PS220)}{card(PS219)} = \frac{3}{4} = 0.75.$$

Let us evaluate the final similarity between two places *PS219* and *PS220*. Previously, we have $\mu_g(PS219, PS220) \approx 0.83$ and $\mu_t(PS219, PS220) = 0.75$. Hence, their final similarity, with explicitly given $w_g = 0.3$ and $w_t = 0.7$, is:

$$\begin{aligned}\mu(PS219, PS220) &= w_g * \mu_g(PS219, PS220) + w_t * \mu_t(PS219, PS220)\\ &= 0.3 * 0.83 + 0.7 * 0.75 = 0.774\end{aligned}$$

Our method are obviously in conformity with the three intuitions mentioned before.

7 Evaluation

We selected 20 situation pairs from what? The situation pairs were chosen semi-randomly. In other words, we restricted the choice as follows: All chosen situations should have at least one specialisation in order to allow the information-theoretic similarity measure to work. Furthermore, we ensured that at least some of the pairs would have ancestor/descendant relationships with each other. Give examples.

8 Shortcomings With Respect to Context Data

The similarity measure has to model properties as well as relations and should provide a structured similarity measure. Furthermore, the approaches must support similarity at a conceptual level, e.g. we do not compare single instances of geometric features, but conceptual descriptions of feature types.

Geometric models use dimensional properties to specify semantics. Relations can only be represented as “compound dimensions”. Geometric models do not allow for structured representation of concepts, involving, for example, hierarchic or part-whole relations, nor for representing the scope of properties, e.g., properties referring only to a part of a concept [31]. Traditional geometric similarity measures do not include relations [20].

For feature models, two individuals cannot be related in a structured way. No partial match would be detected between *Paris* and *France*. Like geometric models, feature models do not support a structured representation of concepts [31].

The strength of the network approach is the representation of relations between concepts. Pure network models do not describe concepts any further (e.g., by features) though they can be combined with feature models such as the semantic neighbourhoods in the Matching-Distance Similarity Model [32].

Alignment models describe the structure of individuals by specifying their properties and the relations between their parts. It provides good results if the individuals compared have an analogous structure (one-to-one mapping and parallel connectivity), but it cannot cope well with inconsistent, complex structures which likely occur in context data [20].

The main problem with transformational models is the lack of a suitably generic set of transformations for geospatial concepts. Therefore, this similarity measure is not yet applicable in practice.

9 Context Data

Measured scales of features (properties) can be ordinal (point) dimension, interval dimension, ratio dimension, binary dimension, categorial dimension, linear dimension, circular dimension. In many cases, the same dimension can be measured with different scales.

Relations describe the connection between two concepts, between a concept and its property or between two properties: Is-a: a concept inherits the additional properties of superconcepts. Belongs-to: Contained- within: Next-to: Part-of: concept and its part are modelled as separate concepts related via part-of relation.

The specification of relations starts with selecting the relevant relations and the concepts the relation refers to. After specifying the direction, weights can be chosen to determine the importance of the relation. Then, the related concept needs to be described within its conceptual space. The related concept again may have relations to other concepts. We call all concepts, directly or indirectly related to a concept C, the semantic neighbourhood of C. A user decides on the size of a semantic neighbourhood.

1. Identify all common and corresponding quality dimensions of query concept q and the compared concept c, as well the dimensions of concepts related to c and q. All additional properties and relations of concept c that the query concept q does not contain are ignored in the similarity measurement.
2. For each related concept crel and qrel compute a semantic distance value analogous to the similarity measure of conceptual spaces.
3. Compute the difference between c and q.

9.1 Point Dimensions

Use the Minkowski metrics for similarity calculation (e.g., the Euclidean metric sums up the squared distances and extracts its root:

$$d(Q,C) = \left(\sum_{i=1}^{n} |q_i - c_i|^r \right)^{1/r},$$

$r = 1$ results in the City-block distance and $r = 2$ in the Euclidian distance. Similarity is then a decaying function of semantic distance $d(Q,C)$.

9.2 Interval Dimension

Maximum similarity is reached when a concept agrees on the exact interval with the query. Any deviation reduces the similarity. Q is the query concept and C is the compared concept (i.e., how similar Q to C). The formula is proposed by [20]:

$$d(Q,C) = \begin{cases} l(Q) - l(Q \cap C) + l(Q \mid C) & \text{if } l(Q \cap C) \neq 0 \\ l(Q - C) & \text{if } l(Q \cap C) = 0 \end{cases}$$

Maximum similarity is reached when a concept has the same or a smaller interval. It is enough that concepts fulfill only a subsection of the query interval.

$$d(Q,C) = \begin{cases} l(C) - l(Q \cap C) & \text{if } l(Q \cap C) \neq 0 \\ l(Q - C) - l(Q) & \text{if } l(Q \cap C) = 0 \end{cases}$$

Maximum similarity is reached when a concept covers at least the same interval as the query.

$$d(Q,C) = \begin{cases} l(Q) - l(Q \cap C) & \text{if } l(Q \cap C) \neq 0 \\ l(Q - C) - l(C) & \text{if } l(Q \cap C) = 0 \end{cases}$$

Maximum similarity is reached when a concept is described by at least one value that is part of the query interval (one of interval). Only non-overlapping intervals affect similarity negatively.

$$d(Q,C) = \begin{cases} 0 & \text{if } l(Q \cap C) \neq 0 \\ l(Q - C) - l(Q) - l(C) & \text{if } l(Q \cap C) = 0 \end{cases}$$

To measure the overall spatial distance between concepts c and q the distance of each property is summed up according to the slightly modified formula of the Minkowski metrics. Q_i is the value on dimension i of query concept q and C_i for concept c.

10 Continuous Data

10.1 Minkowski Distance

It is used to compute distance between two multivariate points. In particular, the Minkowski Distance of order 1 (*Manhattan*) and order 2 (*Euclidean*) are the two most widely used distance measures.

11 Categorical Data

Categorical data is also known as nominal or qualitative multi-state data. The key characteristic of categorical data is that the different values that a categorical attribute takes are not inherently ordered. Thus, it is not possible to directly compare two different categorical values.

For the sake of notation, consider a categorical data set D containing N objects, defined over a set of d categorical attributes where A_k denotes the k^{th} attribute. Let the attribute A_k take n_k values in the given data set that are denoted by the set A_k. We also use the following notation:

- $f_k(x)$: The number of times attribute A_k takes the value x in the data set D. Note that if $x \notin A_k, f_k(x) = 0$;
- $\widehat{p}_k(x)$: The sample probability of attribute A_k to take the value x in the data set D. The sample probability is given by
$$\widehat{p}_k(x) = \frac{f_k(x)}{N}.$$
- $p_k^2(x)$: Another probability estimate of attribute A_k to take the value x in a given data set, given by
$$p_k^2(x) = \frac{f_k(x)(f_k(x)-1)}{N(N-1)}$$

Almost all similarity measures assign a similarity value between two data instances X and Y belonging to the data set D as follows:

$$S(X,Y) = \sum_{i=1}^{d} w_k S_k(X_k, Y_k),$$

where $S_k(X_k, Y_k)$ is the per-attribute similarity between two values for the categorical attribute A_k. Note that $X_k, Y_k \in A_k$. The quantity w_k denotes the weight assigned to the attribute A_k.

11.1 Overlap Measure

The simplest way to find similarity between two categorical objects is to assign a similarity of 1 if the attribute values are identical and a similarity of 0 if the attribute values are not identical. Hence, for two multivariate categorical objects, the similarity between them will be directly proportional to the number of attributes in which they match. This simple measure is also known as the *overlap* measure.

One obvious drawback of the *overlap* measure is that it does not distinguish between the different values taken by an attribute. All matches, as well as mismatches, are treated as equal.

11.2 *Goodall*1

$$S_k(X_k, Y_k) = \begin{cases} 1 - \sum\limits_{q \in Q} p_k^2(q) & \text{if } X_k = Y_k \\ 0 & \text{otherwise} \end{cases},$$

where $Q \subseteq A_k : \forall q \in Q, p_k(q) \leq p_k(X_k)$.

11.3 How to automatically determine importance weights of context attributes

- *Number of values taken by each attribute*, $n_k = |A_k|$: A data set might contain attributes that take several values (e.g., several hundred) and attributes that take very few values. A similarity measure might give more importance to the second attribute, while ignoring the first one.
- *Distribution of* $f_k(x)$. This refers to the distribution of frequency of values taken by an attribute in the given data set. In certain data sets, an attribute might be distributed uniformly over the set A_k, while in others the distribution might be skewed. A similarity measure might give more importance to attribute values that occur rarely, while another similarity measure might give more importance to frequently occurring attribute values.

12 Related Work

Context Matching matches two context data: provided context and desired context. Provided context information is coming from the sensors, other applications, and generally from context providers. On the other hand, desired context information is the query of the context consumers in active or passive format. Matching operation on context data highly depends on the used context model and representation. Use and selection of matching methods are directly related with the representation and defined ontology for context.

Semantic similarity is central to many cognitive processes and plays an important role in the way humans process and reason about information. Information retrieval systems use similarity to detect relevant information for a given query. Current information retrieval systems apply mainly syntactic techniques to determine similarity.

Categorical

Continuous

Hierarchical

Kashyap and Sheth [33] represent a context lattice and its operations such as specificity relationship, overlap, coherent, GLB (the most general context or greatest lower bound).

Jeh and Widom [16] propose a similarity measure called SimRank which claims that two objects are similar if they are related to similar objects.

Anagnostopoulos et al. [28] considered three axioms in terms of similarity measuring: (i) the *disjoint* axiom, (ii) the *closure* axiom and (iii) the *compatibility* relation among concepts. The first axiom denotes that two concepts, A and B, are, by definition, disjoint, i.e., $A \subseteq \neg B$. The closure axiom defines whether existential ($\exists$) and universal ($\forall$) restrictions are applied over a relation/property. The compatibility relation defines whenever two concepts, regardless their disjointness, are characterised compatible or not with respect to their semantic interpretation. Then, they distinguished two types of similarity: Taxonomical Similarity and Relational Similarity. The taxonomical similarity measure of two concepts D and C belong to the same taxonomy H is defined as:

$$sim(D, C) = \frac{|H_C \cap H_D|}{|H_C \cap H_D| + \alpha|H_D \setminus H_C| + \beta|H_C \setminus H_D|}, \tag{7}$$

where $|H|$ is the cardinality of the H set. The factors α, β refer to the weights for common and different features respectively and $\alpha + \beta = 1$. If two concepts are disjoint, then, it is inappropriate to measure their taxonomical similarity.

The relational similarity (RS) between c and d is based on the similarity of their associated concepts with respect to their common relations. Common relations mean that c and d have the same relations, but not necessarily the same ranges.

Tavakolifard et al. [34] propose that two contexts A, B are similar if they have similar aspects. Two aspects c, d are similar if they belong to similar contexts:

$$s(A, B) = \frac{C_1}{|O_A||O_B|} \sum_{i=1}^{O_A} \sum_{j=1}^{O_B} s(O_A^i, O_B^j)$$

C1 is a constant between 0 and 1. OA is aspects of A. OB is aspects of B.

$$s(c, d) = \frac{C_2}{|I_c||I_d|} \sum_{i=1}^{I_c} \sum_{j=1}^{I_d} s(I_c^i, I_d^j)$$

C2 is a constant between 0 and 1. Ic is contexts of c. Id is contexts of d.

[20] propose a hybrid approach for semantic similarity measurement, which can represent the complex semantics of spatial data. It allows for retrieving relevant data by determining the similarity between the query and the semantic descriptions of geographic feature types within the database. The hybrid similarity measure combines the geometric structure of conceptual spaces with the relational structure of semantic nets to one, cognitively plausible knowledge representation with an inherent similarity measure.

13 Conclusion

Computing similarity between categorical attributes has been discussed in a variety of contexts. In this paper, we have brought together several such measures and evaluated them in the context of situation similarity and recognition.

References

[1] A. K. Dey. Understanding and using context. *Personal and Ubiquitous Computing*, 5(1):4–7, February 2001.

[2] C. Bettini, O. Brdiczka, K. Henricksen, J. Indulska, D. Nicklas, A. Ranganathan, and D. Riboni. A survey of context modelling and reasoning techniques. *Pervasive and Mobile Computing*, In Press, Corrected Proof, 2009. ISSN 1574-1192. doi: DOI: 10.1016/j.pmcj.2009.06.002.

[3] S. Loke. On representing situations for context-aware pervasive computing: six ways to tell if you are in a meeting. In *Proceedings of PerCom Workshops*, 2006.

[4] J. Barwise and J. Perry. Situations and attitudes. *The Journal of Philosophy*, 78 (11):668–691, 1981.

[5] Anand Ranganathan and Roy H. Campbell. An infrastructure for context-awareness based on first order logic. *Personal Ubiquitous Comput.*, 7(6):353–364, 2003. ISSN 1617-4909.

[6] N. Goodman. *Seven strictures on similarity*, pages 437–450. Bobbs Merrill, Indianapolis, 1972.

[7] D. Medin, R. Goldstone, and D. Gentner. Respects for similarity. *Psychological Review*, 100(2):254–278, 1993.

[8] Yi Guan, Xiaolong Wang, and Qiang Wang. A new measurement of systematic similarity. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 38(4): 743–758, 2008.

[9] Dekang Lin. An information-theoretic definition of similarity. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1-55860-556-8.

[10] A. Tversky. Features of similarity. *Psychol. Rev.*, 84(4):327–352, 1977.

[11] A. Tversky and I. Gati. *Cognition and Categorization*, chapter Studies of similarity, pages 79–98. Lawrence Erlbaum, Hillsdale, NJ, Septemper 1978.

[12] J. H. Lee, M. H. Kim, and Y. J. Lee. Information retrieval based on conceptual distance in is-a hierarchies. *Journal of Documentation*, 49(2):188–207, 1993.

[13] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on systems, man, and cybernetics*, 19(1):17–30, 1989.

[14] A. Rodriguez. *Assessing semantic similarity among spatial entity classes.* PhD thesis, University of Maine, 2000.

[15] A. Rodriguez and M. J. Egenhofer. Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 15(2):442–456, 2003.

[16] Glen Jeh and Jennifer Widom. Simrank: a measure of structural-context similarity. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543, New York, NY, USA, 2002. ACM.

[17] A.B. Markman and D. Gentner. Structural alignment during similarity comparisons. *Cognitive Psychology*, 25(3):431–467, 1993.

[18] U. Hahn, N. Chater, and L. B. C. Richardson. Similarity as transformation. *Cognition*, 87(1):132, 2003.

[19] K. Janowicz. Sim-dl: Towards a semantic similarity measurement theory for the description logic alcnr in geographic information retrieval. In R. Meersman, Z. Tari, and P. Herrero, editors, *2nd international workshop on semantic-based geographical information systems (SeBGIS06)*, volume 4278, pages 1681–1692, Montpellier, France, 2006. Springer.

[20] Angela Schwering and Werner Kuhn. A hybrid semantic similarity measure for spatial information retrieval. *An Interdisciplinary Journal Spatial Cognition & Computation*, 9(1):30–63, 2009.

[21] Stephen S. Yau and Junwei Liu. Hierarchical situation modeling and reasoning for pervasive computing. In *SEUS-WCCIA '06: Proceedings of the The Fourth IEEE Workshop on Software Technologies for Future Embedded and Ubiquitous Systems, and the Second International Workshop on Collaborative Computing, Integration, and Assurance (SEUS-WCCIA'06)*, pages 5–10, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2560-1. doi: http://dx.doi.org/10.1109/SEUS-WCCIA.2006.25.

[22] Mieczyslaw M. Kokar, Christopher J. Matheus, and Kenneth Baclawski. Ontology-based situation awareness. *Information Fusion*, 10(1):83 – 98, 2009. ISSN 1566-2535. doi: DOI: 10.1016/j.inffus.2007.01.004. URL `http://www.sciencedirect.com/science/article/B6W76-4N3GFMB-1/2/4509936bd7e228b4` Special Issue on High-level Information Fusion and Situation Awareness.

[23] Abraham Bernstein, Esther Kaufmann, Christoph Brki, and Mark Klein. *Wirtschaftsinformatik 2005*, chapter How Similar Is It? Towards Personalized Similarity Measures in Ontologies, pages 1347–1366. Physica-Verlag HD, 2005.

[24] P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

[25] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of 14th International Joint Conference on Artificial Intelligence*, pages 448–453, Montreal, 1995.

[26] Juha Makkonen, Helena Ahonen-Myka, and Marko Salmenkivi. Topic detection and tracking with spatio-temporal evidence. In *Proceedings of the 25th European Conference on IR Research, ECIR 2003*, pages 549–549, Pisa, Italy, April 2003. Springer.

[27] A. B. Kocaballıand A. Koçyiğit. Granular best match algorithm for context-aware computing systems. *J. Syst. Softw.*, 80(12):2015–2024, 2007. ISSN 0164-1212. doi: http://dx.doi.org/10.1016/j.jss.2007.03.006.

[28] C. B. Anagnostopoulos, Y. Ntarladimas, and S. Hadjiefthymiades. Reasoning about situation similarity. In *3rd International IEEE Conference on Intelligent Systems*, pages 109 –114, Sept. 2006.

[29] W. Qin, D. Zhang, Y. Shi, and K. Du. Combining user profiles and situation contexts for spontaneous service provision in smart assistive environments. In *UIC '08: Proceedings of the 5th international conference on Ubiquitous Intelligence and Computing*, pages 187–200, Berlin, Heidelberg, 2008. Springer-Verlag.

[30] S. W. Loke. Representing and reasoning with situations for context-aware pervasive computing: a logic programming perspective. *Knowl. Eng. Rev.*, 19(3): 213–233, 2004. ISSN 0269-8889.

[31] A. Schwering. Hybrid model for semantic similarity measurement. In R. Meersman and Z. Tari, editors, *4th international conference on ontologies, databases, and applications of semantics (ODBASE05)*, volume 3761, pages 1449–1465, Agia Napa, Cyprus, 2005. Springer.

[32] A. Rodriguez and M. J. Egenhofer. Comparing geospatial entity classes: An asymmetric and context-dependent similarity measure. *International Journal of Geographical Information Science*, 18(3):229–256, 2004.

[33] Vipul Kashyap and Amit Sheth. Semantic and schematic similarities between database objects: a context-based approach. *The VLDB Journal*, 5(4):276–304, December 1996. URL `http://dx.doi.org/10.1007/s007780050029`.

[34] Mozhgan Tavakolifard, Svein Johan Knapskog, and Peter Herrmann. Trust transferability among similar contexts. In *Q2SWinet '08: Proceedings of the 4th ACM symposium on QoS and security for wireless and mobile networks*, pages 91–97, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-237-5.