The centroid or consensus of a set of objects with qualitative attributes

Adolfo Guzman-Arenas, Alma-Delia Cuevas, Adriana Jimenez
a.guzman@acm.org, almadeliacuevas@gmail.com, dyidyia@yahoo.com

Centro de Investigación en Computación (CIC) and *Escuela Superior de Cómputo (ESC), Instituto Politécnico Nacional (IPN), México*

ABSTRACT. It is clear how to compute the average of a set of numeric values; thus, handling inconsistent measurements is possible. Recently, using *confusion*, we showed a new way to compute the consensus (a kind of average) of a set of assertions about a non-numeric fact, such as the religion of John.

This paper solves the same problem for a set of *objects* possessing several symbolic attributes. Suppose there is a murder, and we ask several observers about the height, sex, hair color and ethnicity of the killer. They report divergent observations. What is the most likely portrayal of the assassin? Given a bag of assertions about an object described by qualitative features, this paper tells how to assess the most plausible or "consensus" object description. It is the most likely description to be true, given the available information. It is the "centroid" of the bag. We also compute the *inconsistency* of the bag: how far apart the testimonies in the bag are. All observers are equally credible, so differences arise from perception errors, and from the limited accuracy of the individual findings.

Keywords: I.2.4 Knowledge representation; qualitative values; inconsistency; confusion; consensus; truth discovery.

1. Previous work

When measurements on the same quantitative attribute disagree, we resort to the average (or centroid) and variance of the results. We know how to take into account contradicting facts like these, and we do not regard them necessarily as inconsistent. We just assume that the measurers' gauges have different precisions or accuracies. It could also be that observers have a propensity to lie, and in this case we apply the Theory of Evidence (Dempster 1968, Shafer 1979). Or we could use Fuzzy Logic, selecting some sets as possible answers and assigning a degree of membership to each measurement for each set.

(Yin, Han & Fu 2008) provide a manner to find the most likely "truth" among a set of qualitative information[1] obtained from "information providers" in the Web. The information is an assertion about a qualitative value, a "fact" as found in the Web. This work resorts to the "trustworthiness" of each informant (resembling Dempster-Schafer), as well as a measure of the similarity among two of these non-numeric values (resembling our *confusion*, as defined in next pages).

A recent paper (Guzman-Arenas & Jiménez, *to appear*) finds the centroid or most likely value of a bag of qualitative values, such as {Afghanistan; Beirut; Iraq; Kabul; Middle

[1] Qualitative attributes (such as *religion* or *hair color*) are also called non-numeric properties, aspects, features, or linguistic variables. The values these attribute attain (such as *Muslim* or *brown*) are called qualitative values, non numeric values, or linguistic constants.

East; Afghanistan; Syria}. The answer is not necessarily the most popular value or mode (Afghanistan), nor the least common ancestor (Middle East). The answer is not based on the probability that informants lie (like in the theory of evidence), nor it contains fuzzy values. The answer assumes that all informants are equally credible, and the discrepancy of their findings arises from the way or method used when obtaining their observations.

As an example, let us assume that we want to discover what pet Bart has, so we ask several observers to find out. One of them hears the animal bark, so he reports "a dog;" another observer finds fur hairs, so she reports "a mammal," while a third observer reports "a large dog," seeing the silhouette of the animal at night. Assume the reported values are {dog; mammal; German Shepherd; iguana}. One of them is the most likely pet. If we select "dog," reporter 1 is happy (he shows no discomfort, since our selection agrees with his report, a dog); reporter 2 is also happy (our selection agrees with her finding, a mammal); reporter 3 is somewhat displeased, since he observed a more accurate dog (a German Shepherd), not just "some dog." Reporter 4 is more uncomfortable, since he found an iguana. If our selection is "iguana", only reporter 4 is at comfort, while three others are somewhat upset. If we could measure these discomforts, we could select as the most likely pet (consensus value) *the value that minimizes the sum of disagreements* for all the observers when they learn of the value chosen as the consensus value.

The discomfort or disagreement when value *r* is reported instead of the "true" value *s* (as found by the observer) is called the *confusion* in using *r* instead of *s* (Levachkine, Guzman & de Gyves 2005, Levachkine & Guzman 2007). To measure this, it is necessary to give all observers the same *context*, that is, the same set of possible qualitative answers as well as how these are related by specificity or generality. This set is called a *hierarchy* (Figure 1); it is a tree where each node is a qualitative value or, if it is a set, then its immediate descendants form a *partition* of it.

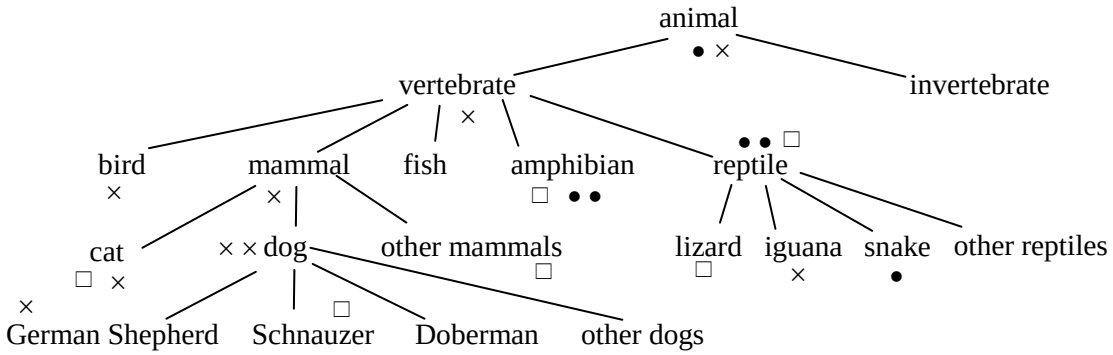

Figure 1. A hierarchy of symbolic values is a tree where every node is either a symbolic value or, if it is a set, then its immediate descendants form a partition. Hierarchies make possible to compute the confusion conf(*r*, *s*) that results when value *r* is used instead of *s*, the true or intended value. The confusion (§1.1) is the number of *descending* links in the path from *r* to *s*, divided by the height of the hierarchy. For instance, conf(dog, Doberman) = 1/4, conf(Doberman, dog) = 0, conf (Doberman, German Shepherd) = 1/4, conf (Doberman, iguana) = 2/4, conf(iguana, Doberman) = 3/4. Observe that conf ∈ [0, 1]. Refer to Section 1.1. The values marked ×, □ and • are used in examples 5 and 6 of Section 1.2

Using hierarchies, next section (§1.1) tells how to compute the confusion among two qualitative values, while section 1.2 explains how to find the consensus or most likely value of a bag of qualitative values.

Sections 1.1 and 1.2 report some previous work, necessary to understand this article. Our contributions appear in section 2.1, which describes how to find the confusion when using an object *O* instead of the real or intended object *O*', and in section 2.2, which obtains the consensus, centroid or most plausible object in a bag of objects, as well as the inconsistency of such bag, a number that reveals how disparate are its members.

1.1 Measuring the confusion between two qualitative values

Work on *confusion* has been reported elsewhere [4, 5, 9, 10]; this section is placed here for completeness, in order to understand Section 2. How close are two numeric values v_1 and v_2? The answer is $|v_2 - v_1|$. How close are two symbolic values such as *cat* and *dog*? The answer comes in a variety of similarity measures and distance functions. Hierarchies (introduced in Figure 1) allow us to define the confusion conf(*r, s*) between two symbolic values. The function conf will open the way to evaluate in Section 1.2 the inconsistency of a bag of symbolic observations. We assume that the observers of a given fact (such as *the killer*) share a set of common vocabulary, best arranged in a hierarchy. A hierarchy can be regarded as the "common terminology"[2] for the observers of a bag: their *context.* Observers reporting in other bag may share a different context, that is, another hierarchy.

What is the capital of Germany? *Berlin* is the correct answer; *Frankfurt* is a close miss, *Madrid* a fair error, and *sausage* a gross error. What is closer to a *cat*, a *dog* or an *orange*? Can we measure these errors and similarities? Can we retrieve objects in a database that are close to a desired item? Yes, because qualitative variables take symbolic values such as *cat, orange, California, Africa*, which can be organized in a hierarchy *H*, a mathematical construct among these values. Over *H*, we can define the function *confusion* resulting when using a symbolic value instead of another.

Definition. For *r, s* ∈ *H*, the **absolute confusion** in using *r* instead of *s*, is

CONF(*r, r*) = CONF(*r*, any ascendant of *r*) = 0;

CONF(*r, s*) = 1 + CONF(*r*, father_of(*s*)).

To measure CONF, count the descending links from *r* (the replacing value) to *s* (the intended or real value). CONF is neither a distance nor an ultradistance function.

We can normalize CONF by dividing it into *h*, the height of *H* (the number of links from the root of *H* to the farthest element of *H*), yielding the following

Definition. The **confusion** in using *r* instead of *s* is

conf(*r, s*) = CONF(*r, s*)/*h*.

Notice that $0 \le \text{conf}(r, s) \le 1$. It is not symmetric: $\text{conf}(r, s) \ne \text{conf}(s, r)$, in general. The function conf is not a distance function, but it obeys the triangle inequality [6].

Example 1. For the hierarchy of Figure 1, CONF(cat, mammal) =0; if I ask for a mammal and I am given a cat instead, I am happy, and CONF=0. But CONF(mammal, cat) =1; if I ask for a cat and I get a mammal, I am somewhat unhappy, and CONF=1. For the same

[2] If the symbolic values become full *concepts,* it is best to use an *ontology* instead of a *hierarchy* to place them [2].

reason, CONF(cat, vertebrate)=2. Being given a vertebrate when I ask for a cat makes me unhappier than when I was handed a mammal.

Example 2. In the hierarchy of Figure 1, conf(cat, dog) =1/4; conf(cat, Schnauzer) =1/2.

Remark. Since symbolic values lie in a hierarchy, it is not possible for a value to have two immediate ascendants, to have more than one path from it towards the root. That is, *rabbit* may not be both a mammal and a bird.

The type of hierarchy of Figure 1 is the most common type, and it is sometimes called a *normal* hierarchy, as opposed to ordered (§1.1.1) or percentage hierarchies (§1.1.2).

1.1.1 When symbolic values are totally ordered

If the values in a hierarchy could be ordered by a "<" relation, for instance cold < chilly < tepid < warm < hot < burning, we will have an *ordered hierarchy* [5], with height 1 (Figure 2) always. The confusion for ordered hierarchies with *n* children is:

conf (*r*, *r*) = *conf* (*any child*, *root*) = 0;

conf (*root*, *any child*) = 1;

conf (*r*, *s*) = (number of steps to go from *r* to *s*) / (*n*-1) when *r* and *s* are children.

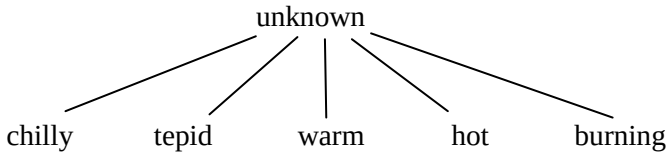

Figure 2. An ordered hierarchy about temperatures, with root *unknown* and five children, which are ordered by a "<" relation. conf (*unknown*, *chilly*)=1; conf(*warm*, *unknown*)=0; conf(*chilly*, *tepid*) = conf(*tepid*, *chilly*) = conf(*tepid*, *warm*)=1/4; conf(*chilly*, *warm*) =1/2; conf(*hot*, *chilly*)=¾; conf (*chilly*, *burning*)=1

Example 3. For hierarchy (B) of Figure 6, conf(short, unknown) =conf(medium, unknown) =conf(tall, unknown) =0; conf(unknown, short) =conf(unknown, medium) =conf(unknown, tall) =1; conf(short, medium) =conf(medium, short) =conf(medium, tall) =conf(tall, medium) =½; conf(short, tall) =conf(tall, short) =1.

1.1.2 When the sizes of the sets of the hierarchy are known

If the number of elements of the sets forming a hierarchy is known, we have *percentage hierarchies* [9]. For instance, consider the countries from the American Continent (Figure 3), also called America (not to be confused with USA, a country in America). If I ask for a North American person and they give me a Mexican, the confusion is 0, since Mexicans are North Americans. If I ask for a Mexican and they give me a North American, the confusion is 1-(100/500), since in a total of 500 million people, only 100 million are Mexicans. Thus, for percentage hierarchies, the confusion is

conf(*r*, *r*) = conf(*r*, *s*) = 0 when *r* is any descendant of *s*;

conf(*r*, *s*) = 1 – relative proportion of *s* in *r*, when *s* is a descendant of *r*;

conf(*r*, *s*) = 0 + conf(father_of(*r*), *s*), otherwise.

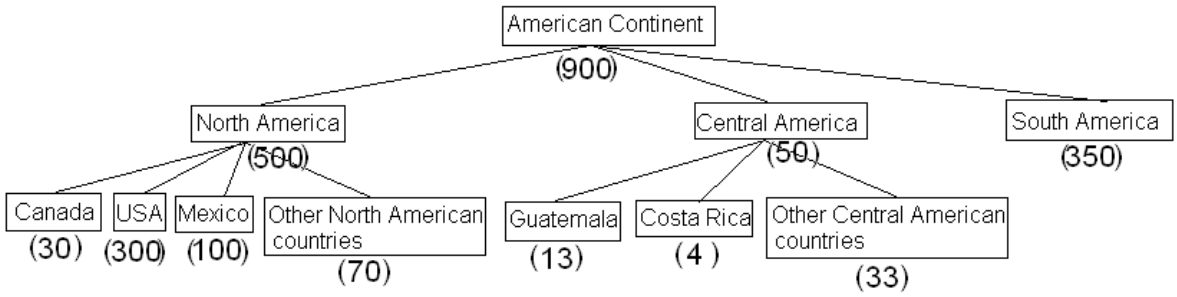

Figure 3. Hierarchy of the American Continent, with the number of inhabitants in millions. conf(*South American*¸ *American*)=0; conf(*American*, *South American*)=1-(350/900). conf(*Canadian*, *Mexican*) =0 + conf(*North American*, *Mexican*)=0+1-(100/500)=4/5. conf(*Canadian*, *Costa Rican*) =1-(4/900). Notice that conf never reaches 1, unless some set is empty. In this example, the largest confusion is conf(*American*, *Costa Rican*). Notice that for *America* we mean the American Continent, not the USA

Example 4. For the hierarchy in Figure 4, the confusion in using a pitcher instead of a catcher is conf(pitcher, catcher) =8/9, whereas the confusion in using a pitcher instead of a base player is conf(pitcher, base player) =2/3. See Table 1.

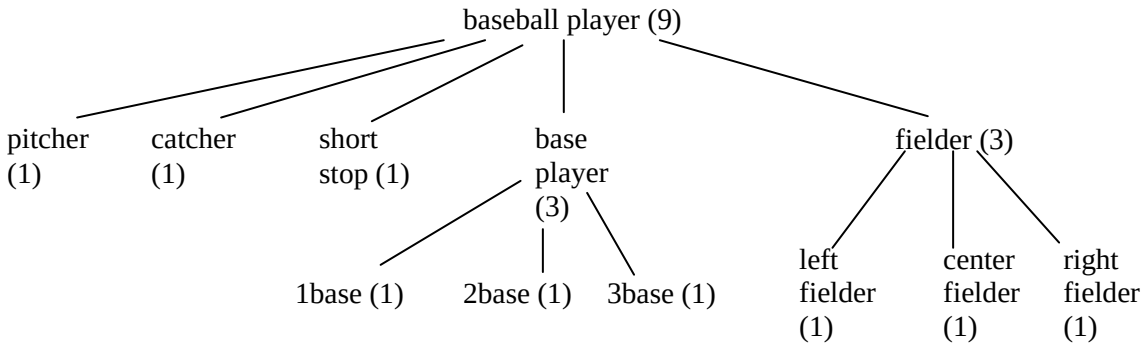

Figure 4. A percentage hierarchy is one where the sizes of the sets are known. The confusion of using a left fielder when I want a fielder is 0, whereas the confusion of using a fielder instead of a left fielder is 2/3

1.1.3 Properties, uses and comments about conf

The theoretical properties and uses of conf are covered elsewhere [4, 5, 9, 10]; nevertheless, we distill some remarks here.

Finding a good (or a better) metric for gauging the similarity of qualitative data. There are many proposed similarity functions already for qualitative data, for instance the Jaccard correlation. Is *conf* a better function? While it is easy to agree on the difference between 7 and 13, there is no "better" way to agree on the distance between, say, "dog" and "cattle." We are now 6.6×10^9 human beings; no matter what metric we design, there will be some subset of people which do not find it useful, meaningful, or pertinent. It is better to classify these similarity functions according to their intended use, that is, to the *context* in which they will be exploited. In this sense, *conf* is useful for problems where qualitative data have shades of details (different precisions, or "granularity"), "Doberman" being more precise than "dog" and much more than "mammal." Notice also that

"mammal" can be made more precise in several ways, one leading to "Doberman", another leading to "sperm whale", for instance.

What definition of *conf* is better? The appropriate definition depends on the context or intended application. Thus, we offer several definitions (sections 1.1, 1.1.1, 1.1.2 and 3.1.2), but concentrate on one: the normalized confusion conf(*r*, *s*) = CONF(*r*, *s*)/*h*.[3]

Since *conf* is not a distance, is it a bad similarity function? The idea that the similarity between two qualitative values *should be a distance* is debatable. We offer *conf* which distinguishes between the confusion produced when I want a dog and they give me a Doberman, and the confusion produced when I want a Doberman and they give me a dog. This is no "better" or "worse" than other functions –it depends on its use, *its context.* We believe it is more appropriate in the presence of degrees of detail (degrees of granularity, of accuracy).

What are the mathematical properties of *conf*? These are given elsewhere [4, 5, 10].

How practical is *conf* to use? How useful is it? Has it been used in some real examples? *conf* is easy to implement and use; it has been used to solve some practical problems, see [9, 10]. A forthcoming paper will tell how *conf* is used to *cluster* qualitative data.

Table 1. For Figure 4, the table shows conf(*r*, *s*) when using the value of row *r* instead of the intended value in column *s*. For instance, conf(short stop, baseball player) = 0; conf(short stop, catcher) =8/9; conf(short stop, fielder) = 2/3. For clarity, empty boxes have value 8/9.

	bbp	**pit**	**cat**	**ss**	**base**	**fielder**	**1b**	**2b**	**3b**	**LF**	**CF**	**RF**
baseball player (bbp)	0				2/3	2/3						
Pitcher (pit)	0	0			2/3	2/3						
Catcher	0		0		2/3	2/3						
short stop (ss)	0			0	2/3	2/3						
base player (base)	0				0	2/3	2/3	2/3	2/3			
fielder (outfielder)	0				2/3	0				2/3	2/3	2/3
first base (1b)	0				0	2/3	0	2/3	2/3			
second base (2b)	0				0	2/3	2/3	0	2/3			
third base (3b)	0				0	2/3	2/3	2/3	0			
left fielder (LF)	0				2/3	0				0	2/3	2/3
center fielder (CF)	0				2/3	0				2/3	0	2/3
right fielder (RF)	0				2/3	0				2/3	2/3	0

1.2 Measuring the degree of inconsistency of a bag of qualitative values

The *inconsistency* of a bag of values is reported elsewhere [6]; this section is placed here for completeness, in order to understand Section 2.

The setting is that several observers report (qualitative) values about a given property of an object they were asked to observe. These values –a bag of them– may be different, but our observers are not liars (so that the theory of evidence does not apply). Their reported values are crisp (no fuzzy values are reported –no fuzzy logic needs to be used). The explanation for not everybody reporting the same value is that the way they observed ("meas-

[3] Equally important is the *hierarchy* employed –this sets the context of the observations.

ured" or gauged) the property was different –their methods of observation had different precision; accuracy varies.

Problem 1. Given a bag of observations reporting non-numeric values, how can we measure its inconsistency? What is the value that minimizes this inconsistency? We shall call r^* this value and σ the inconsistency that r^* produces. Notice that *inconsistency* is a property of a bag of values, not of a single value.

Restrictions to the solution to Problem 1:

(A) All the reported values are about the same *fact* or property. One observer can not report about the identity of the killer, while another observer tells us about the weather in London.

(B) The fact or feature that the observers are gauging, has a single value. There is only one killer. The weather in London (for a particular date and corner of the city) is unique.

(C) All reporters use the same *context* expressed in a vocabulary arranged in the same hierarchy –the same hierarchy for all observations in a bag. It contains all possible answers. It is clear that for observers with conceptions about the animals (and their differences) disagreeing with those of Figure 1, the consensus r^* will differ. Thus, r^* and σ are a function of the bag and the universe of possible values (the hierarchy).

Intuitively, r^* is the value most likely to be true, given the available information, and taking into account observation errors. One of the values of the bag must be the most plausible value, the consensus. Since all observers are equally believable, we could find the confusion of any given observer with respect to any reported value r --a kind of "discomfort" measured by conf(r, s) when value r is preferred or reported, instead of the value s reported by him/her. Adding[4] these confusions for all observers, we find the total confusion (total "discomfort") that such value r produced (if it were selected as the "consensus") in all observers. There must be a value r^* that produces the lowest total confusion. Such r^* is the consensus or centroid of the bag. The inconsistency of the bag, called σ, is such minimum divided by the number of elements of the bag. Thus, we have

Solution. The *centroid* or *consensus* r^* of a bag B of observations reporting qualitative values $\{s_1, s_2, \ldots, s_n\}$ is the value $r_j \in B$ that minimizes

$$\sum_{i=1}^{n} \text{conf}(r_j, s_i) \qquad \text{for } j = 1,..., n \tag{1}$$

For each r_j, this sum is the total confusion that r_j produces among all elements of the bag. The *inconsistency* σ of B is the minimum that such r^* produces, divided by n:

$$\sigma = (1/n) \min_{j\in[1,n]} \sum_{i=1}^{n} \text{conf}(r_j, s_i) = (1/n) \sum_{i=1}^{n} \text{conf}(r^*, s_i) \tag{2}$$

[4] Weights could be given to observers if we think they have different precisions in their observations. We don't do this, for simplicity.

Example 5. For bag_3= {air, airplane, land, road, subway, subway, motorcycle}, marked with × in Figure 5, the total confusion for air is 0 +1/3 +2/3 +1/3 +3/3 +3/3 +3/3 =4.333; for airplane, is (0+0+1+2+3+3+3)/3 =4; for land, it is 3.333; for road is 2.666; for subway is 2; for motorcycle is 2.333. Thus, its consensus is subway, and its inconsistency is 2/7. *Example 6.* For bag_1 = {animal, vertebrate, bird, mammal, cat, dog, dog, iguana, German Shepherd}, marked with × in Figure 1, its consensus or centroid *r** is German Shepherd, and its inconsistency is [(0+0+1+0+1+0+0+2+0)/4]/9 =1/9. *Example 7.* For bag_2 = {animal, amphibian, amphibian, reptile, reptile, snake}, marked with • in Figure 1, *r** = snake, σ = 1/12. *Example 8.* For observations with □, *r** = Schnauzer, σ = (6/4)/6 = ¼.

1.2.1 Properties, uses and comments about centroid and inconsistency

The inconsistency σ and centroid *r** of a collection of values, as well as their theoretical properties, comparisons with similar measures and uses, are dealt elsewhere [6]; for that reason, we only list here a short list of their properties.

I. σ and *r** are the solutions to Problem 1.
II. σ and *r** are properties of the bag, and depend on the context of use –represented by the hierarchy employed. The role of the hierarchy in the solution to Problem 1 is to provide a *common vocabulary* for all observations. See restriction 1.2.(C).
III. The inconsistency $\sigma \in [0, 1)$. In fact, for a bag *B* of size *n*, $0 \leq \sigma \leq (n-1)/n$.
IV. There may be more than one value *r** that minimizes the total confusion.
V. To compute the inconsistency of a bag, we resort to finding *r** first. In other words, the inconsistency of a bag is the average total discomfort (average total confusion) produced by *r**. This is the lowest discomfort attainable; any other element different from *r** will give a larger or equal total confusion (by definition of *r**).
VI. The consensus *r** is not inevitably the most popular value (the mode), which is dog in Example 6 for the elements marked with (×), while *r** =German Shepherd. Also, the consensus is not inevitably the most precise (deeper in the hierarchy) value in the bag: a repeated less precise observation in the bag may become the centroid.
VII. The *least common ancestor* (vertebrate in example 7) in general is not the centroid, since it produces a total confusion larger or at best equal than the total confusion produced by *r**. It is "too general" for many of the observations.
VIII. Given the consensus *r** of *B*, there is no $r' \in B$ such that *r'* is a descendant of *r**.
IX. The total confusion induces a total ordering on the elements of a bag; that element with the lowest total confusion is called the *centroid.*

Notice that we have found a way of adding (and averaging) apples and oranges, and a quantity (σ) to quantify out how disperse or divergent a bag of symbolic values is. Also, the appropriate *conf* to compute *r** and σ depends on the problem at hand (Cf. §1.1.3.).

2. Inconsistency of a bag of objects

Sometimes, observations come in bundles; that is, several properties of the same object are observed by a reporter, returning a list like (tall, airplane, Mexico) as the result of her observation, meaning perhaps that the person was tall, traveled by airplane, and lives in Mexico. Several observation lists reported by several observers about the same object may

show discrepancies; thus, it is sensible to compute the inconsistency of a bag of observation lists. For this, we must find first a way to find the discomfort or confusion when object *O* is used instead of *O*'.

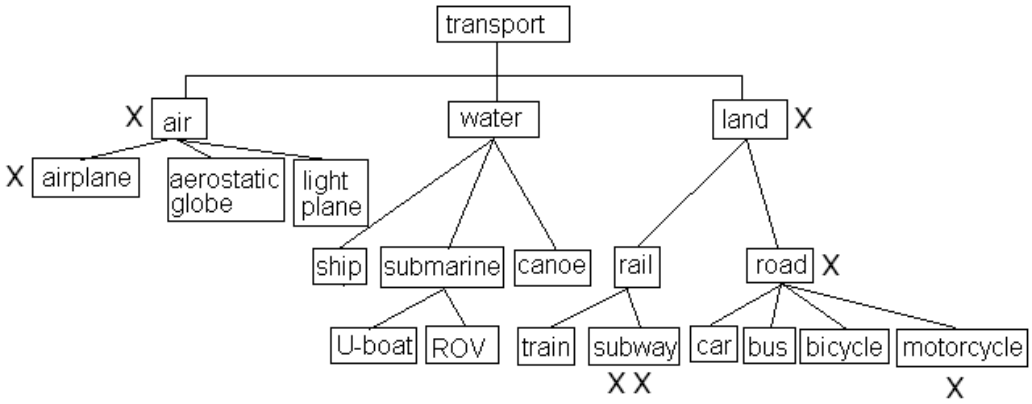

Figure 5. A hierarchy of types of transportation. The observations of bag_3 (see example 5 above) are shown with ×. conf(airplane, transport) =0; conf(transport, airplane) =2/3; conf(U-boat, ROV) =conf(U-boat, canoe) =1/3; conf(U-boat, motorcycle)=1

2.1 Object comparison

For us, an object is described by an indivisible list of values of some selected relevant properties or *aspects*. (See footnote 1.) Thus, we may say *O* = (short, Cuba, iguana), meaning perhaps that person *O* is short, lives in Cuba and has an iguana as pet.

When comparing two objects, we must observe:

A. Objects can be compared only if they have the same set of attributes. An object with attributes religion, ethnicity and age can not be compared against another object with attributes pet, ethnicity and height.
B. Each attribute is single-valued. Thus, an object can not have two ages.
C. For every attribute, all reporters use the same *context* or vocabulary, arranged in the same hierarchy –the same hierarchy for all observations of an attribute. [5] Different attributes (such as ethnicity and religion) will use different hierarchies, in general.

Notice the similarity of these restrictions with restrictions (A) to (C) of §1.2.

In what follows, objects have inseparable properties which are not transferable –they can not be independently considered when comparing objects, nor can they be transferred "freely" from an object to another (detached and remixed). For instance, we observe at some distance two persons coming down from an airplane. One of them seems to us to be a tall woman, oriental looking, and had long hair; the man was short, black and with short hair. We did not see the long hair on the black man, and we did not see a short oriental lady. This is true irrespective of whether the properties of an object are *statistically independent* or not: it is true irrespective of the fact that height is correlated with gender. Sec-

[5] What is a reasonable context to use? We can suppress improbable values from our hierarchy (values unlikely to be reported in the situation at hand). But: (a) the hierarchy must obey the partition property (See caption to Figure 1); so, introduce suitable nodes such as "Other mammals", "Other reptiles," in appropriate places; (b) if you shorten the height of the hierarchy, you increase the confusion among values –more discerning power; (c) do not eliminate improbable nodes if you want to have an accurate estimate of the importance of confusing a value with another.

tion 3.1.1 treats the case where properties of an object *can* be mixed with other object's properties, as in the case "I saw somebody with a gun, but I am not sure whether it was the oriental woman or the black man." A forthcoming paper computes the confusion for objects that have statistically dependent properties. Thus, in what follows, properties of an object are statistically independent and non-detachable.

2.1.1 Confusion between two objects

If O_1 =(tall, Mexico, iguana) and O_2 =(tall, American Continent, reptile), then we can measure the total confusion of using O_1 instead of O_2 by a weighted addition[6] of the confusions that their respective properties provoke, thus: conf(O_1, O_2) =$w_1$conf(tall, tall) + $w_2$conf(Mexico, American Continent) + $w_3$conf(iguana, reptile), using the hierarchies of Figure 1 and Figure 6. For simplicity, we set all the weights to be 1 (the reader may assign them other values according to the problem at hand). Then, conf(O_1, O_2) =0 +0 +0 =0, whereas conf(O_2, O_1) =conf(tall, tall) +conf(American Continent, Mexico) +conf(reptile, iguana) =0 +2/3 +¼ =0.92,

To obtain a normalized confusion between two objects, we divide the above sums by their number of properties. For the example, conf(O_2, O_1) = 0.92/3 = 0.31. Thus, we extend conf to work on objects described by *m* (non-detachable, statistically independent) properties, as follows:

Definition. The confusion conf(*O, O*') obtained when object *O* is used instead of *O*', is

$$\text{conf}(O, O') = (1/m) \sum_{i=1}^{m} \text{conf}(o_i, o_i') \qquad \text{where } o_i \ (o_i') \text{ is the } i\text{th property of } O \ (O').$$

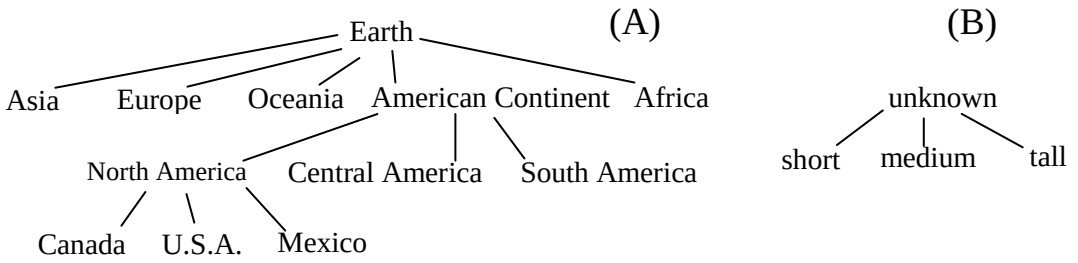

Figure 6. Hierarchies for places to live. (A), a normal hierarchy, and height of persons (B), an ordered hierarchy. conf(Mexico, North America) =0; conf(North America, Mexico) =1/3; conf(Mexico, South America) =2/3; conf(South America, Mexico)=1. conf(short, unknown) =1; conf(unknown, short) = conf(unknown, medium) =1; conf(short, medium) =conf(medium, tall) =1/2; conf(short, tall) =1

Example 9. I want a base player that has a North American wife. Thus, using hierarchies of Figure 3 and Figure 4, *O* = (base, North America). I am given instead a third base player that has a wife living in the American Continent, *O*' = (3b, American Continent). My confusion when given *O*' instead of the desired person *O* is conf(*O*', *O*) = [conf(3b, base) +conf(American Continent, North America)]/2 = [0 +(1-(500/900))]/2 =2/9. If I were given instead *O*'' = (3b, Mexico) my confusion conf(*O*'', *O*) would be [conf(3b, base)

[6] The weights w_i reflect the relative importance of properties –those with larger weights are more reliable to observe. For simplicity, we omit the weights, setting them all equal to 1, thus considering all properties equally observable.

+conf(Mexico, North America)]/3 = [0+0]/3 =0. If I were given a player *O*''' = (2b, Canada) my confusion will still be [0+0]/3 =0. Notice that both players *O*'' and *O*''' fulfill my wish: a base player with a wife from North America.

Example 10. Six observers were sent to evaluate the properties *height*, *place of living* and *pet* of the same person, and were given as vocabulary the hierarchies of Figure 1 and Figure 6. Their reports appear in Table 2. The confusion obtained by using observation *i* instead of observation *j* appears at the intersection of row *i* and column *j* in the right part of the table. For instance, conf(O_3, O_2) = [1 + 0 + 1/4]/3 =0.417

Table 2. *Left side:* The object reported by each observer appears in a row. *Right side:* It represents the confusion when object in row *i* is used instead of object in column *j*. Hierarchies of Figure 1 and Figure 6 are used. Thus, conf(O_1, O_3) =[conf(tall, short) + conf(Mexico, Canada) + conf(iguana, Schnauzer)]/3 =[1+1/3+3/4]/3 =0.694. Arithmetic is done with 8 decimals; results shown are rounded to three.

Observer	height	Place of living	Pet	1	2	3	4	5	6
1	tall	Mexico	iguana	0	0	0.694	0.167	0.361	0
2	tall	American Continent	reptile	0.306	0	0.806	0.167	0.361	0
3	short	Canada	Schnauzer	0.611	0.417	0	0.167	0.278	0
4	medium	American Continent	vertebrate	0.556	0.25	0.639	0	0.194	0
5	medium	Africa	mammal	0.667	0.361	0.667	0.111	0	0
6	unknown	Earth	animal	0.917	0.611	1	0.528	0.611	0

2.2 Consensus (most likely object) and inconsistency of a bag of objects

How to select, from a bag {O_1, O_2,…O_n} of *n* observed objects,[7] the most likely or most plausible object? As we did for bags of properties, we can compute the total confusion that object O_1 produces in the bag when it is selected as "the answer." The total confusion for O_1 is just conf(O_1, O_1) +conf(O_1, O_2) +…+conf(O_1, O_n). We could compute also the total confusion for O_2, and for every object in the bag. It makes sense to select that object that has the lowest total confusion as *the consensus* of the bag. No other object will produce lower total confusion. Then, we can define the inconsistency of the bag as that lowest total confusion divided by *n*.

Therefore, we can find the inconsistency and centroid of a bag of *objects* using the same formulas of §1.2 "Measuring the degree of inconsistency of a bag of qualitative values," but using in them conf for objects (§2.1.1).

*Example 11. F*or bag {1,2,2,3} of objects in Table 2, the consensus is object 1 with an inconsistency of [conf(1,1) +conf(1,2) +conf(1,2) +conf(1,3)]/4 =(0+0+0+0.694)/4 =0.174. [8]

*Example 12. F*or bag {2, 3, 4, 4, 5}, the consensus is object 3 with an inconsistency of (0.417 +0 +0.167 + 0.167 +0.278)/5 =0.206

*Example 13. F*or bag {2, 4, 6}, the consensus is object 2 with an inconsistency of (0 + 0.167 + 0)/3 =0.056.

The following formulas formalize the results.

[7] Remember: the object is the same; it just appears different to different observers, due to the way they obtained their observations.

[8] Total discomforts are: for object 1, 0+0+0+0.694=0.694; for object 2 is 0.306+0+0+0.806=1.111; for object 3 is 0.611+0.417+0.417+0 =1.445. Thus, object 1 produces the lowest total discomfort (total confusion); therefore, object 1 is the centroid or consensus of the objects (observations) of the bag {1, 2, 2, 3}.

The *centroid* or *consensus* O^* of a bag B of objects $\{O_1, O_2, \ldots, O_n\}$ described by qualitative values, is the object $O_j \in B$ that minimizes

$$\sum_{i=1}^{n} \text{conf}(O_j, O_i) \qquad \text{for } j = 1,..., n$$

For each O_j, this sum is the total confusion that O_j produces among all objects of the bag.

The *inconsistency* of the bag is the minimum that such O^* produces, divided by n:

$$\sigma = (1/n) \min_{j\in[1,n]} \sum_{i=1}^{n} \text{conf}(O_j, O_i) = (1/n) \sum_{i=1}^{n} \text{conf}(O^*, O_i)$$

The objects in the bag are all described by the same properties or attributes (such as place of origin, color of hair, religion…); the *values* of such properties will vary, of course, from object to object. A Ph. D. thesis (Jiménez, *in press*) contains more details.

Remarks. For a bag of objects,

I. O^* and σ depend on the context of use –represented by the hierarchies employed. The role played by the hierarchies is to provide a *common vocabulary* for all observations.
II. The inconsistency $\sigma \in [0, 1)$. In fact, for a bag of size n, $0 \le \sigma \le (n\text{-}1)/n$.
III. There may be more than one value O^* that minimizes the total confusion.
IV. To compute the inconsistency of a bag, we resort to finding O^* first. O^* produces the lowest total confusion attainable for the bag. Its inconsistency is such lowest total confusion divided by the size of the bag.
V. The consensus of a bag is not necessarily the *mode* or most popular object of such bag. For instance, the *mode* for bag {1,2,2,3} is object 2, but its consensus is object 1.
VI. Observed values can not be "detached" from the objects. That is, values of any object come together. They come "bundled." In other words, the second observation (a tall person) of Table 2 owned a reptile. No observation reported a tall person owning a Schnauzer, although Schnauzer was one of the pets reported. There were no observations reporting a short person living in Africa or owning an iguana. The implication is that we can not compute the centroid of the observed heights, the centroid of the observed places of living, and the centroid of the observed pets, and report as the centroid of the bag the object having these three centroids as values. To witness, for bag {2,3,4,4,5} we have:
Centroid of heights =centroid of {tall, short, medium, medium, medium} =medium;
Centroid of places =centroid of {American Continent, Canada, American Continent, American Continent, Africa} =Canada;
Centroid of pets =centroid of {reptile, Schnauzer, vertebrate, vertebrate, mammal} = Schnauzer;
But the object possessing as values these centroids, that is, the object (medium, Canada, Schnauzer) is *not* the consensus of bag {2,3,4,4,5}. It is not even in the bag! It was not observed. The correct value is object 3 = (short, Canada, Schnauzer).

VII. "General" or vague observations such as observation 6 = (unknown, Earth, animal) will provoke large discomforts among other observations (see last row of Table 2). If they take part in a bag of observations, they will unlikely be selected as the consensus, since other, more specific observations, have "something more to say," have more information.

VIII. The function "total confusion among objects" induces a total order in a bag of objects, $O_1 \leq O_2 \leq O_3 \leq \ldots \leq O_n$. The "smallest" object, O_1, is the *centroid* of the bag.

Example 14. What do we know about Mexican composer Agustin Lara? What is his correct name? Where was he born? Some people assert that in Madrid, while most persons believe he was born somewhere in Mexico. He was born at the end of the 19th century or beginning of the twentieth century. Therefore, the context for our problem is given by hierarchies (A), (B) and (C) of Figure 7. After consulting the Web, some results are:

a. According to ServicioWeb.cl, his name is Agustin Lara (http://www.servicioweb.cl/bolero/agustin_lara.htm), and he was born in Mexico City on October 30, 1896 .
b. Wikipedia (http://es.wikipedia.org/wiki/Agust%C3%ADn_Lara) tells us that he was born in Tlacotalpan, October 30, 1900 whereas his name is Agustín Lara y Aguirre del Pino.
c. For redescolar.ilce: Agustin Lara; October 30, 1897; in D.F. http://www.redescolar.ilce.edu.mx/redescolar/publicaciones/publi_quepaso/agustin_lara.htm.
d. For BioStars International (http://www.biosstars-mx.com/a/agustin_lara.html): the country of Mexico, October 30, 1897, the name Ángel Agustín María Carlos Fausto Mariano Alfonso del Sagrado Corazón de Jesús Lara y Aguirre del Pino.
e. The page last.fm (http://www.lastfm.es/music/Agustin+Lara) reports Agustin Lara, Tlacotalpan, October 30, 1900.

Table 3 shows these findings, while Table 4 shows the confusion conf(R, S) when object R is used instead of object S. For instance, the confusion when observation **a** is used instead of observation **c** is [conf(Agustin Lara Aguirre del Pino, Agustin Lara Aguirre del Pino) +conf(Mexico City, D.F.) +conf(1896, 1897)]/3 =[0 + 0 +1/2]/3 =0.167.

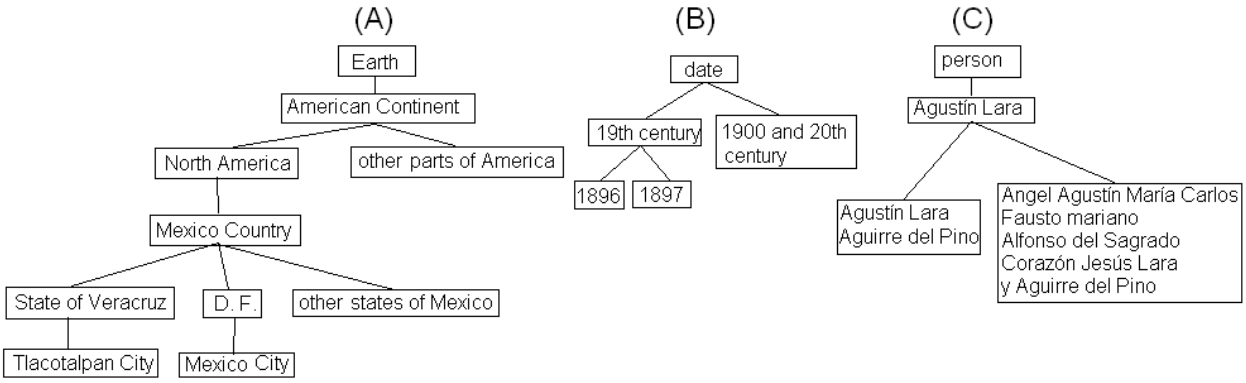

Figure 7. Possible places of birth (A); possible birth dates (B), and possible names (C) (as obtained after the experiment) for Agustin Lara. So as not to confuse it with Mexico City, we call "Mexico Country" to Mexico

Table 3. Five Web pages report different values for attributes of composer Agustin Lara.

Observer	Name of composer	Place of birth	Date of birth
a. servicio.web	Agustín Lara Aguirre del Pino	Mexico City	30 oct 1896
b. Wikipedia	Agustín Lara	Tlacotalpan	30 oct 1900
c. redescolar	Agustín Lara Aguirre del Pino	D.F.	30 oct 1897
d. BioStars	Ángel Agustín María Carlos Fausto Mariano Alfonso del Sagrado Corazón de Jesús Lara y Aguirre del Pino	Mexico Country	30 oct 1897
e. last.fm	Agustín Lara	Tlacotalpan	30 oct 1900

Table 4, for Example 14. The confusion conf(*R*, *S*) when object *R* is used instead of object *S*, is in the intersection of row *R* and column *S*. For instance, the confusion between **e** and **d** is [conf(Agustin Lara, Ángel Agustín María Carlos Fausto Mariano Alfonso del Sagrado Corazón de Jesús Lara y Aguirre del Pino) +conf(Tlacotalpan, Mexico Country) + conf(30 October 1900, 30 October 1896)]/3 = [1/2 +0 +1]/3 =0.5.

Observation	a	b	c	d	e
a	0.0	0.3	0.167	0.333	0.3
b	0.633	0.0	0.567	0.5	0
c	0.233	0.3	0.0	0.167	0.3
d	0.467	0.3	0.233	0.0	0.3
e	0.633	0.0	0.567	0.5	0.0

The confusion when object **a** is used instead of **b** is [conf(Agustin Lara Aguirre del Pino, Agustin Lara) +conf(Mexico City, Tlacotalpan) + conf(30 October 1896, 30 October 1900)]/3 = [0 +2/5 +1/2]/3 =0.3.

From the five observations, we can now find the most likely value for Agustin Lara. If we believe all five pages, that is, for bag {**a**,**b**,**c**,**d**,**e**} we have: the total confusion when **a** is used as candidate for the consensus of {**a**,**b**,**c**,**d**,**e**} is 0 +0.3 +0.167 +0.333 +0.3 =1.1; when using **b**, it becomes is 0.7633 +0 +0.567 +0.5 +0 =1.7; when using **c** as candidate, the total confusion is 0.233 +0.3 +0 +0.167 +0.3 =1.0; using **d**, it becomes 0.467 +0.3 +0.233 +0 +0.3 =1.3; while if **e** is used, it is 0.633 +0 +0.567 +0.5 +0 =1.7. The observation producing the lowest total confusion is **c.** *Thus, the consensus of bag* {**a**,**b**,**c**,**d**,**e**} *is composer* **c** *and its inconsistency is 1.0/5 =0.2.*

But notice that inconsistency is a property of *the bag* of observations, not of a single observation. Thus, if for some reason we discard observation **d** (perhaps because it is the greatest outlier [6], the "most far away" report), we have, for bag {**a**,**b**,**c**,**e**}: the total confusion when **a** is used as candidate for the consensus of {**a**,**b**,**c**,**e**} is 0 +0.3 +0.167 +0.3 =0.767; when using **b**, it is 0.633 +0 +0.567 +0 =1.2; using **c** is used as candidate, the total confusion is 0.233 +0.3 +0 +0.3 =0.833. Finally, using **e**, it is 0.633 +0 +0.567 +0 =1.2. Therefore, for bag {**a**,**b**,**c**,**e**}, the observation producing the lowest total confusion (lowest total discomfort) is **a.** *Thus, the consensus of bag* {**a**,**b**,**c**,**e**} *is* **a** *and the inconsistency of the bag is 0.767/4 =0.192.* [9]

For the bag {**a**,**c**,**e**}, **a** produces a total discomfort of 0.467; **c** produces 0.533; **e** produces 1.2. Hence, for that bag *the consensus is observation* **a** *and the inconsistency of the bag is 0.156.*

[9] The inconsistency of bag {**a**, **b**, **c**, **e**} is lower than that of {**a**, **b**. **c**, **d**, **e**}, since we discarded **d**, the greatest outlier.

For bag {**a**,**b**,**b**,**c**,**c**,**d**,**e**}, the consensus is object **c** and the bag's inconsistency is 0.186.

Example 15. Ten observers (which could be considered as witnesses of an event happening in Montana, USA, about police action) watched, for eight seconds, the video in http://www.espacioblog.com/notaroja/post/2006/06/27/el-video-policiaco-la-semana. We then handed them the hierarchies of Figures 8, 9 and 10, asking them the questions:

- In what vehicle the action took place?
- What was its color?
- How many policemen took part?
- What was the ethnicity of the persecuted person?
- What was the final action? How did the event ended?

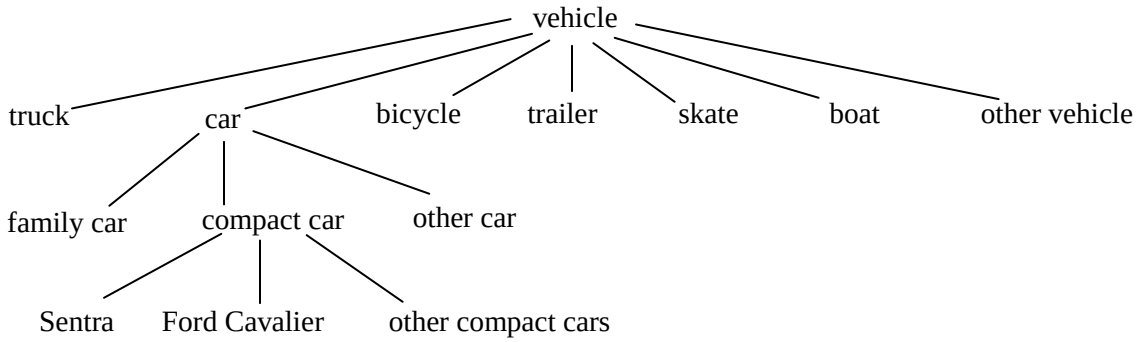

Figure 8. Hierarchy about vehicle types. The vehicle seen by the eyewitnesses of Example 15 was of one of these types

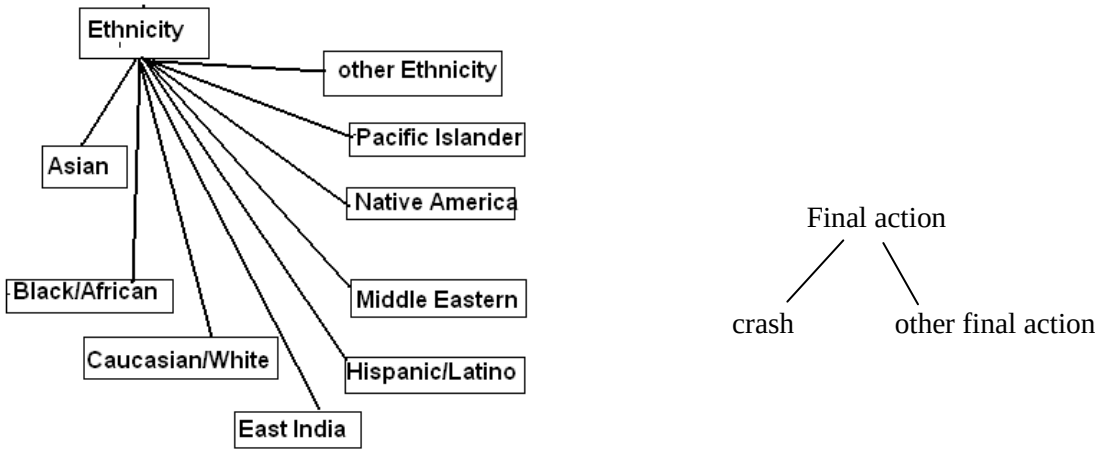

Figure 9. The context (a hierarchy) about the ethnicity of the perpetrator, and the final action hierarchy, of the success witnessed in example 15

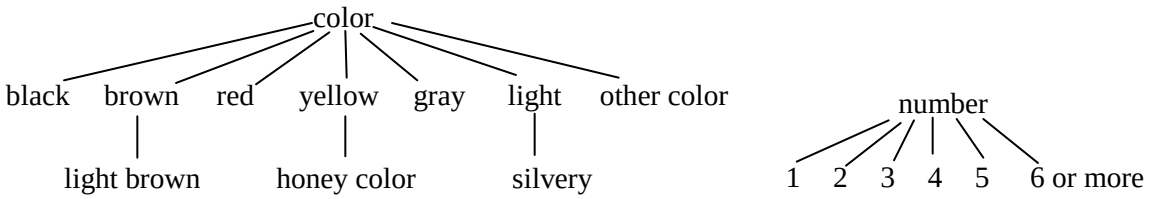

Figure 10. Hierarchies about color and about the number of policemen taking part in the event, as perceived by witnesses of Table 5

The witnesses' answers are shown in Table 5.

Table 5. Answers obtained by ten witnesses of the police persecution of Example 15.

Obs	vehicle	color	number	Ethnicity	Final Action
1	compact car	yellow	2	Caucasian/White	other final action
2	compact car	yellow	3	Caucasian/White	crash
3	sentra	honey color	3	Native American	crash
4	Ford cavalier	light brown	3	Caucasian/White	crash
5	compact car	yellow	3	Caucasian/White	crash
6	compact car	light brown	3	Hispanic/Latino	crash
7	compact car	gray	3	Hispanic/Latino	crash
8	familiar car	light brown	4	Caucasian/White	crash
9	compact car	light brown	3	Caucasian/White	crash
10	compact car	silvery	4	other Ethnicity	crash

Using the formulas for confusion among objects, we obtain the confusions between each par of observations, shown in Table 6.

Table 6. Confusion among observations of the video of Example 15. The confusion conf(*R*, *S*), when object of row *R* is used instead of the intended object of column *S*, is shown in this table. For instance, conf(2, 4) =0.267; conf(2,5) =0. Each object is an observation containing five qualitative values, as reported in Table 5.

Obs	1	2	3	4	5	6	7	8	9	10
1	0.0	0.24	0.607	0.507	0.24	0.64	0.54	0.547	0.44	0.68
2	0.24	0.0	0.367	0.267	0.0	0.4	0.3	0.307	0.2	0.44
3	0.44	0.2	0.0	0.467	0.2	0.4	0.3	0.507	0.4	0.44
4	0.34	0.1	0.467	0.0	0.1	0.2	0.3	0.107	0.0	0.44
5	0.24	0.0	0.367	0.267	0.0	0.4	0.3	0.307	0.2	0.44
6	0.54	0.3	0.467	0.267	0.3	0.0	0.1	0.307	0.2	0.44
7	0.54	0.3	0.467	0.467	0.3	0.2	0.0	0.507	0.4	0.44
8	0.447	0.207	0.573	0.173	0.207	0.307	0.407	0.0	0.107	0.467
9	0.34	0.1	0.467	0.067	0.1	0.2	0.3	0.107	0.0	0.44
10	0.58	0.34	0.507	0.507	0.34	0.44	0.34	0.467	0.44	0.0

Now, if we take into account only observations 1,3,5,7 and 9, the results computed by our program are given in Figure 11. *The object with lowest total discomfort is observation 5 with an inconsistency of 0.221 for bag 1, 3, 5, 7, 9. Thus, for this bag the consensus is observation 5.*

If instead, we use evidence from witnesses 2,4,6,8,10 to solve the case, the results are: observation: 2 produces a total confusion (total discomfort) of 1.413; observation 4 yields 0.847; observation 6 gives 1.313; observation 8 produces 1.153; and observation 10 produces 1.753. *The object with lowest total discomfort is observation 4 with an inconsistency of 0.169 for bag 2,4,6,8,10. Thus, for this bag the consensus is observation 4.*

If we take into account all ten observations, the results are: observation 1 produces a total confusion (total discomfort) of 4.44; observation 2 produces 2.52; observation 3 produces 3.353; observation 4 produces 2.053; observation 5 produces 2.52; observation 6 produces 2.92; observation 7 produces 3.62; observation 8 produces 2.893; observation 9 produces 2.12; and observation 10 produces 3.96. *The object with lowest total discomfort is observation 4 with an inconsistency of 0.205 for bag 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. Thus, for this bag the consensus is observation 4.*

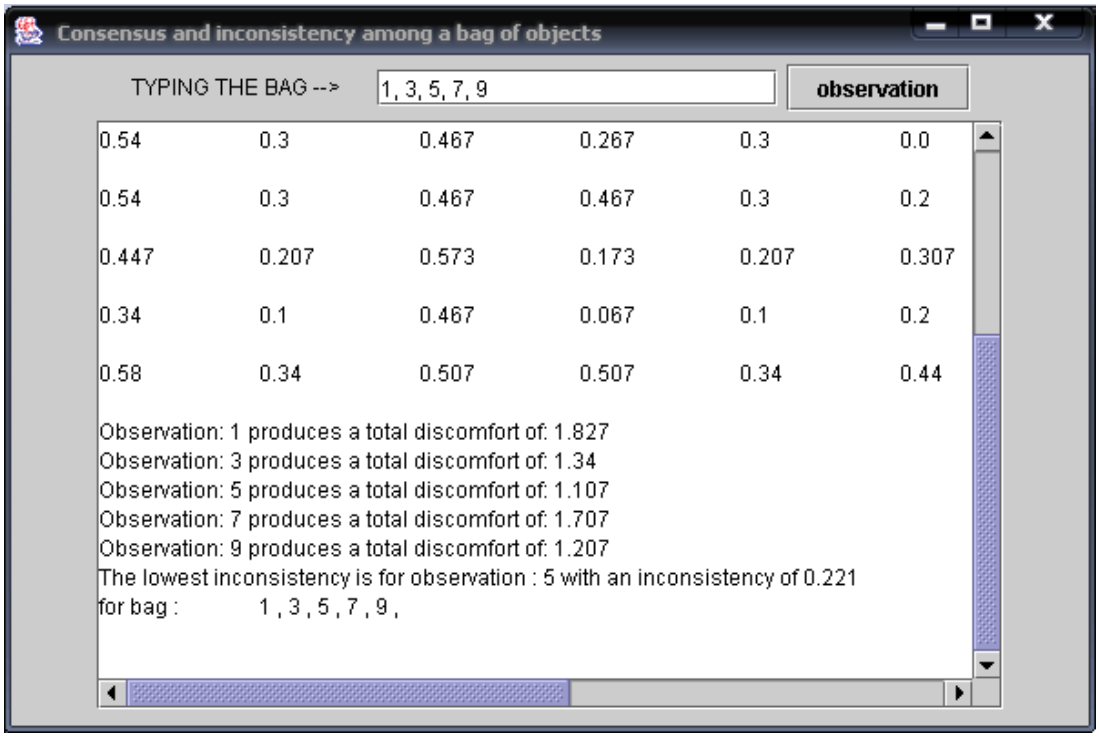

Figure 11. The results of the program when only observations 1, 3, 5, 7, 9 of the video are taken into account. It shows that the consensus is observation 5. The first five rows of results shown are the last rows of Table 6

Example 16. Eighteen observers watched the picture in http://www.premier-ministre.gouv.fr/es/ for 10 sec. Then they were asked: **1)** Who is the person signing the document? **2)** This person traveled to which country? **3)** Who is the person standing? **4)** The person signing was accompanied during his trip by …; **5)** What is the color of the tie of Spain's President? **6)** The person who signs the document is…; **7)** About which topics they chatted? Simultaneously, to help them select the answers, they were given the hierarchies of Figure 12 for questions 1,3,4, and those below for 2,5,6,7, shown in text form to save space:

2) continent { America [HIERARCHY FOR QUESTION 2]
Europe {Spain Sweden}
Africa
Australian Continent
a country which is not in the tree but I do not remember}

5) color { a color that I do not remember {red black white} [HIERARCHY FOR QUESTION 5]
It did not have }

6) I do not remember { left-handed person; right-handed person } [HIERARCHY FOR QUESTION 6]

7) Anything, I do not remember {one of these subjects { communitarian subjects; political subjects;
war subjects } [FOR QUESTION 7]
any subject { the war in Gaza strip; about America } }

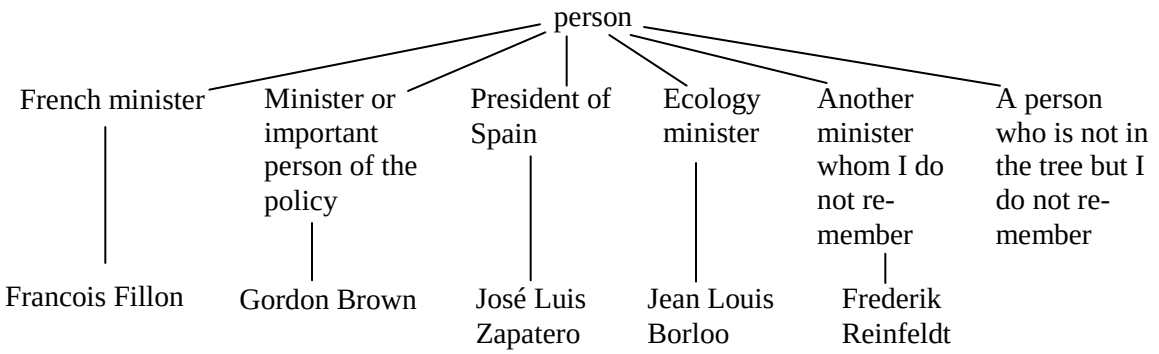

Figure 12. Hierarchy of persons possibly appearing in the signature of an important European document. It is used to answer questions 1, 3 and 4 of Example 16

Their answers appear in Table 7, while Table 8 shows the consensus of several bags.

Table 7. Answers of the 18 observers that watched the signing of an important document of example 16.

Obs.	Who signs?	Traveled to	Standing?	Companion	Tie color	Handed?	Topics
1	Francois Fillon	Spain	Perso not in tree	Perso not in tree	red	Left handed	America
2	Francois Fillon	Spain	J L Zapatero	ecology minister	white	Left handed	Don’t rememb
3	another minister	Spain	president of Spa	person	color	Don't rememb	Don’t rememb
4	another minister	Country not in t	person	person	Don’t rememb	right-handed	one of these su
5	Perso not in tree	Spain	J L Zapatero	J L Borloo	red	Left handed	Political subjec
6	Francois Fillon	Spain	president of Spa	Gordon Brown	red	Don't rememb	one of these su
7	person	Spain	J L Zapatero	Francois Fillon	Don’t rememb	Left handed	one of these su
8	Francois Fillon	America	J L Zapatero	ecology minister	red	right-handed	Political subjec
9	French minister	Spain	Francois Fillon	Perso not in tree	red	right-handed	one of these su
10	Francois Fillon	Spain	person	person	red	Don't rememb	Don’t rememb
11	French minister	Spain	J L Borloo	Gordon Brown	Don’t rememb	Don't rememb	Political subjec
12	Perso not in tree	Spain	J L Zapatero	ecology minister	red	left-handed	military subjec
13	person	Europe	person	President of Spa	red	Don't rememb	any subject
14	person	continent	person	person	white	Don't rememb	Don’t rememb
15	Francois Fillon	Spain	J L Zapatero	Gordon Brown	black	right-handed	one of these su
16	J L Zapatero	Europe	Francois Fillon	person	red	left-handed	political subjec
17	Francois Fillon	Spain	president of Spa	Gordon Brown	Don’t rememb	Don’t rememb	Don’t rememb
18	person	Spain	president of Spa	person	Don’t rememb	right-handed	Don’t rememb

Table 8. Depending on which testimonies are used to gauge the event of July 27, 2007 (example 16), the consensus (center column) varies. In bag 1, all observations are considered. Bag 4 is bag 3 with observation 14 eliminated. From bag 4 to bag 5, observation 5 is expunged. Bag 6 contains the odd observations, while bag 7 contains the even observations. The consensus or centroid (observation 12) of bag 8 is also its mode. That is not the case for bag 9, where its centroid (observation 8) is not its mode.

Bag	Witnesses from Table 7 whose observations are taken into account	Consensus of the bag (most likely object observed)	Inconsistency of bag
1	All 18 observations of Table 7	15	0.222
2	The first 15 observations of Table 7	15	0.229
3	{1,2,5,6,7,8, 9, 10, 11, 12, 13, 14, 15}	5	0.242
4	{1,2,5,6,7,8, 9, 10, 11, 12 , 13, 15}	5	0.256
5	{1,2,5,6,7,8, 9, 10, 11, 12, 15}	15	0.266
6	{1, 3, 5, 7, 9, 11, 13, 15, 17}	15	0.23
7	{2, 4, 6 , 8, 10, 12, 14, 16, 18}	12	0.214
8	{2, 4, 6,6, 8,8, 10, 12,12, 14, 16, 18}	12	0.22
9	{2, 4,4, 6, 8, 10,10, 14,14, 16,16, 18,18}	8	0.203

2.3 The object most consistent with a given predicate P

A related problem is to find how close an object is to a predicate P(*O*).

Let P be a predicate that evaluates to a number between 0 (false) and (1) true; when applied to an object *O*, P(*O*) evaluates the *inconsistency* between *O* and the predicate P.

It is possible to find which of the objects of a set is most consistent with a given predicate P. This problem is solved in [9], where it is called "object *O* fulfils predicate P with confusion ε." Using our definitions, the desired object is that which minimizes P(*O*). Thus, we can talk about the inconsistency *between an object and a predicate.*

3. Conclusions and discussion

Conclusions. Recently [6], using *confusion,* we showed a new way to compute the consensus of a collection of assertions about a non-numeric attribute; that is, the centroid and inconsistency of a bag of symbolic values.

Work reported here is the continuation of that line of work. It solves the same problem for a set of *objects* possessing several symbolic attributes. Given a bag of assertions about an object described by qualitative features, this paper provides a method to assess the most plausible or "consensus" object description. It is the most likely description to be true, given the information available. It also shows how to compute the *inconsistency* of the bag, which measures how far apart the testimonies in the bag are. All observers are equally credible, so differences arise from perception errors, due to the limited accuracy of the individual findings (the limited information extracted by each examination method from the observed reality).

Work herein reported resembles [12], which uses the trustworthiness of informants, which we do not. More over, we determine more precisely the similarity between two qualitative values (using conf). Also, this paper extends conf to objects, and ends up computing the centroid or most plausible value in a bag of objects, as well as the inconsistency of the bag.

3.1 Discussion

Numeric values have a meaning "of their own," so it is easy to agree in the difference between, say, 7 and 13. Opposite to that, symbolic values (such as Mexico City or airplane) must have a *context* against which their closeness or difference can be gauged. This context is provided by hierarchies, over which the function conf is defined. The function conf(*r*, *s*) measures the confusion when symbolic value *r* is used instead of the real, or intended, value *s*. A hierarchy of possible or likely symbolic values is always a parameter of conf. Our definitions of *conf* make it suitable for situations where symbolic values evoke different precisions –for instance, Doberman is more precise than dog, and much more than mammal.

When several numeric measurements are performed over the same property of the same object, and a bag of different values is obtained, how can we measure the inconsistency of the bag?[10] What is the most likely value for the property? Ordinary Logic tells us that there is no such value, and the inconsistency of the bag is 1 (False), since the measurements are incompatible. To compute the most likely value, the Dempster-Schafer theory [3, 11] resorts to the likelihood of different measurers telling the truth (their credibility).. Fuzzy logic can also be used. Other researchers [1, 7] count how many predicates are violated by the bag of observations, and that count is the inconsistency of the bag. For most of us, the most likely value for that property, given a bag of measurements, is just the average or centroid of these values, and the inconsistency of the bag is just the variance σ of the observations.

Our paper [6], summarized in §1.2, provides a way, using conf, to obtain the centroid (*r**) and inconsistency (σ) of a bag of symbolic values reported (by several observers) for the same observed property of the same object. The observers are equally credible, so that their dissimilar observations are due to the difference in their method or instrument of observation. For crude methods, the observed value has "limited precision," while other observers could obtain more detailed measurements (positioned deep in the hierarchy).

For us, an object is represented by a list of qualitative and numeric values. Using *r** and σ for a bag of values, work herein presented makes two contributions:

(a) It obtains the confusion when using object *O* instead of the intended object *O*';

(b) Given a bag of objects as reported by different observers, and using conf(*O*, *O*') for objects, it gives a manner to compute the most likely or most plausible object in such bag. This object (*O**) is called the *consensus* or centroid of the bag. We also compute the inconsistency (σ) of the bag. The objects in the bag are symbolic descriptions of *the same object* in real life –such object is just perceived differently by each observer.

The centroid of a bag of objects can be seen as the "average" of those objects, similar to the average of a bag of numeric measurements. If the bag contains observations about the same object, it is the object most likely to be the real one, given the evidence obtained by the observations. It is the object that produces the lowest total discomfort among all the objects in the bag. Also, the inconsistency (a number between 0 and 1, not just 0 or 1) can be regarded akin to the variance of a bag of numeric measurements.

Will it be feasible in some situations to have two (or more) centroids, for a given bag of objects? For instance, we have *two killers* in an assassination. Or, if we were selecting a

[10] I mean a bag of assertions, such as {the length is 7.2; the length is 7.29; the length is 6.85; the length is 7}.

President, why not have two, not just one? A co-presidency. Total discomfort will be lower, no doubt. Such work is reported in [8].

3.1.1 When the attributes of an object are detachable

Some times, the values observed can be considered as not attached to a particular object,[11] contrary to remark VI of §2.2. For instance, in example 16 (the European politicians), it may be reasonable to assume that the observers saw the colors of the ties, but they can not remember whether a given tie was in the signer or in the standing man. In this case, it is reasonable to apply the algorithm of remark VI: independently find the centroids of each property, and report as the most likely candidate the object having these centroids as values. Beware: the centroid thus found may not in the bag of reported observed objects: it means that the consensus of a bag of observed objects may be an object *not in the bag!*

3.1.2 When a value carries more information than it seems

In general, "dog" means "any dog," including Doberman, Schnauzer and "other dogs" (Figure 1). It is the same situation when we see that "30" means any of 29.6, 29.7…, 30.4. Similarly, "Schnauzer" does not include "dog" (a general dog), since it means "a dog of this particular breed" while "dog" means "a general dog." For this reason CONF(Schnauzer, dog) =0 whereas CONF(dog, Schnauzer) =1. There is more information (more precision) in 30.2 than in 30, as there is more in Schnauzer than in dog.

Nevertheless, at times "dog" may mean "a general dog, not a Doberman specifically nor a Schnauzer specifically nor a …," in the same manner than 30 at times may mean 30.0 (that is, not 30.1 nor 30.2 nor…). We can handle this case by modifying the definition of confusion, taking into account, in the journey from *r* to *s* in the hierarchy, the number of *ascending* links too, not just the descending links (Cf. definition in §1.1), as follows:

CONF(*r*, *r*) = 0;
CONF(*r*, *s*) = max(ascending links from *r* to s, descending links from *r* to *s*) otherwise.

Acknowledgments. Work was in part supported by CONACYT grant 43377 and by SNI.

References

All URLs in this paper were last time consulted February 7, 2009.

1. Byrne, E., & Hunter, A. (2005) Evaluating violations of expectations to find exceptional information, *Data and Knowledge Engineering*, **54**(2):97-120.
2. Cuevas, A and Guzman, A. (2008) A language and algorithm for automatic merging of ontologies. Chapter of the book *Handbook of Ontologies for Business Interaction,* Peter Rittgen, ed. IGI Global (formerly Idea Group Inc.), USA. 381-404
3. Dempster, Arthur P. (1968) A generalization of Bayesian inference, *Journal of the Royal Statistical Society*, Series B, Vol. **30**, pp. 205-247.
4. Guzman, A., and Levachkine, S. (2004) Hierarchies Measuring Qualitative Variables. *Lecture Notes in Computer Science* LNCS **2945**, Springer-Verlag. 262-274. http://www.divshare.com/download/6271736-09c

[11] This is not the same as being statistically independent.

5. Guzman, A., and Levachkine, S. (2004) Graduated errors in approximate queries using hierarchies and ordered sets. *Lecture Notes in Artificial Intelligence* LNAI **2972**, Springer-Verlag. 139-148. ISSN 0302-9743
6. Guzman-Arenas, A., and Jimenez, A. Obtaining the consensus and inconsistency among a set of assertions on a qualitative attribute. *Journal Expert Systems with Applications* **48**(2), to appear.
7. Hunter, A., and Konieczny, S. (2008) Measuring inconsistency through minimal inconsistent sets. *Proc.* 11th *International Conference on Knowledge Representation*, AAAI Press
8. Jimenez, A. (2009) Characterization and measurement of logical properties of qualitative values organized in hierarchies. Ph. D. Thesis. CIC-IPN. In Spanish. In press.
9. Levachkine, S, Guzman-Arenas, A. and de Gyves, V.P. (2005) The semantics of confusion in hierarchies: from theory to practice. In *Contributions to ICCS 05 13th International Conference on Conceptual Structures: common semantics for sharing knowledge*. Kassel, Germany. 94-107 http://www.divshare.com/download/6257877-9a7
10. Levachkine, S. and Guzman-Arenas, A. (2007) Hierarchy as a new data type for qualitative variables. *Journal Expert Systems with Applications* **32**, 3, 899-910.
11. Shafer, Glenn. (1979) A Mathematical Theory of Evidence. Princeton University Press.
12. Xiaoxin Yin, Jiawei Han, Philip S. Fu. (2008) Truth Discovery with multiple conflicting information providers on the Web. *IEEE Trans. KDE* **20**, 6, 796-808.