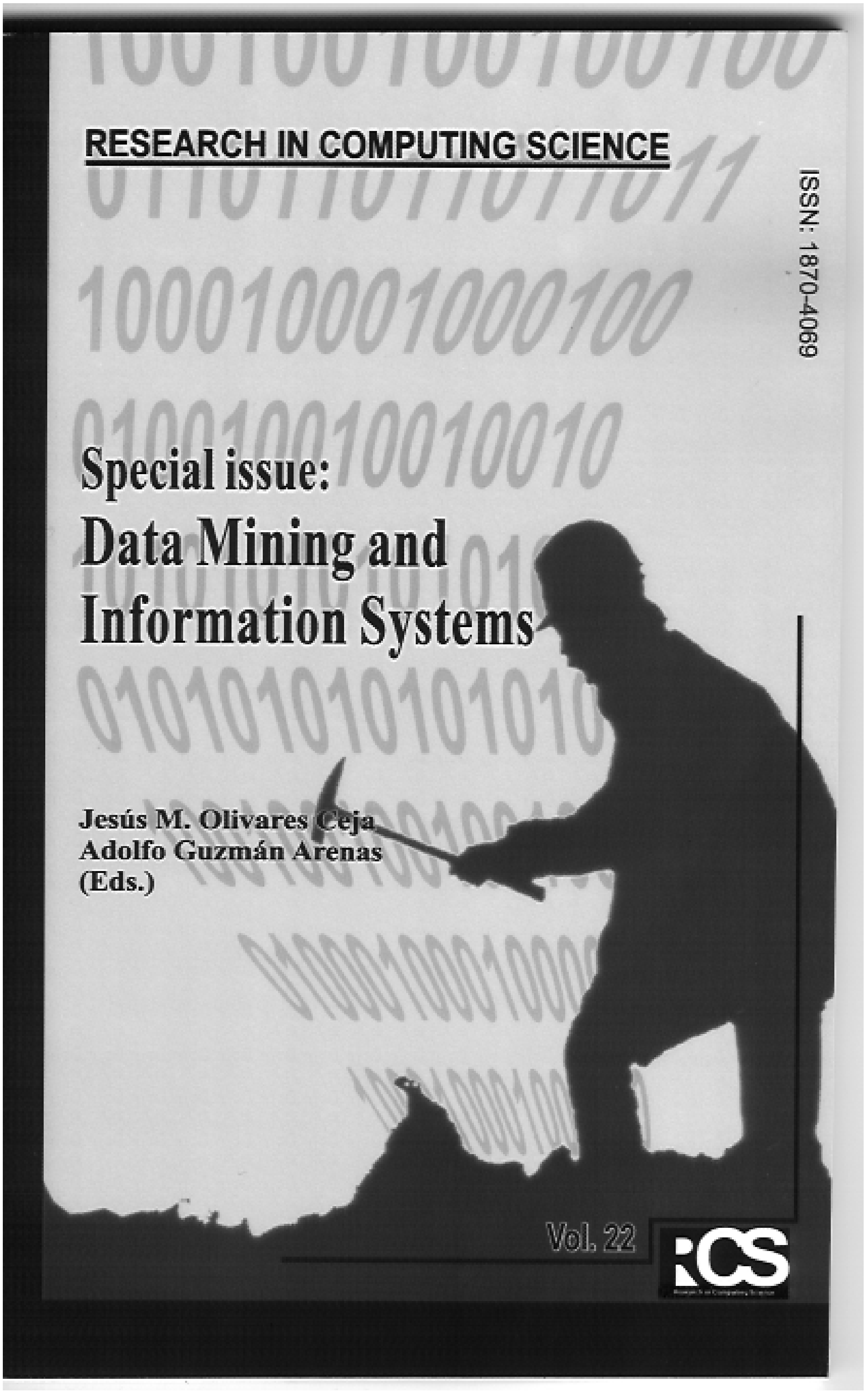


RESEARCH IN COMPUTING SCIENCE

ISSN: 1870-4069

Special issue:
Data Mining and Information Systems

Jesús M. Olivares Ceja
Adolfo Guzmán Arenas
(Eds.)

Vol. 22

RCS

Table of Contents

Indice

Page/Página

Information Systems

Data Mining

Measuring Inconsistency over a Hierarchy of Qualitative Facts

Adolfo Guzmán-Arenas, Adriana Jiménez-Contreras

Centro de Investigación en Computación (CIC),
Instituto Politécnico Nacional (IPN)
07799 Ciudad de México, MÉXICO

a.guzman@acm.org
adrianaj@sagitario.cic.ipn.mx

Abstract. In this article we present a model to compute the degree of inconsistency of a particular event. This situation is described through facts from observers, where each one of them informs on a fact. Contrary to the theory of Dempster-Schafer, all the observers are equally believable since if their observations differ, it will be for those different used observation ways. That is to say, if somebody has the situation in which he wants to investigate and to determine in which transport traveled Luis, and the informants report on what they observed, they will say that he traveled by airplane, bus, train, etc.; then with the model we can specify the most likely fact. We define the way to determine the disagreement of these facts and to determine which will be the value average that adjusts better. It is adjusted to the reported facts.

Keywords: Confusion, inconsistency, fact, observer, center of gravity.

1 Introduction

In this paper we present the study of situations or circumstances of reality, related to a particular aspect, where a serie of consistent or inconsistent observations is exposed as facts that describe the event. For such study we present a model that allows finding the inconsistency in that set of facts.

The data in those facts are qualitative in nature "Jon's hair is black"; more precisely, constants (such as "black") must belong to a hierarchy [6].

The model helps us to determine the degree of inconsistency of each fact in an event, using a function (confusion) in the hierarchy of facts and another function that computes the value of the inconsistency.

The facts are obtained through observers call reporters or informants, these facts can be located over a hierarchy of facts (qualitative values), on this hierarchy, a func-

© Jesús Olivares, Adolfo Guzmán (Eds.)
Data Mining and Information Systems.
Research in Computer Science 22, 2006, pp. 15-27

tion measures the confusion that arises when we use r instead of s, the intended or correct value. For example, about the confusion of using "America" instead of "México"In summary, we study existent facts about an event.

2 Antecedents

Inconsistency is a topic intensely studied in the area of computation. For example, in databases since the integrity is highly appreciated, so that measuring the inconsistency in the data is important.

Also inconsistency in the requirements stage of the development of a system is required, since it is impossible to design a system with inconsistent requirements [3].

Other investigations on this topic are carried out in the analysis of news using Classic logic [7], to find the inconsistency of news over a particular event [1]. Also others have used the Theory of Dempster-Shafer, also known as the Theory of Functions of Beliefs, which is a generalization of the Bayesian theory of subjective probabilities, where the idea is to obtain degrees of beliefs (informants are not reliable, they may lie) for a question and to combine such degrees when they are based on independent elements of evidences. For example, we want to know the probability rained in Mexico City on May 10, 2006; if Juan said that it rained, and Pedro said that it didn't rain. A subjective probability is assigned to the reliability of each people, these events are considered as independent and they combine these degrees of beliefs to determine if it rained or not. This is another form of finding inconsistency in a particular situation [10].

In this paper we solve the following:

- Given an event (set of facts) how certain is it? (To measure the certainty). That is to say, the list of facts will allow us to determine the consistency or inconsistency of these facts, and we will also find the must likely fact, that which generates the smallest uncertainty with respect to all facts in the set.

For example, the color of Luis' hair, an observer says that it is red, others say that it is light brown, light dark, blond and black respectively; it is required to find this set of facts, as well as to determine as close as possible the true color of Luis' hair.

- We want to analyze the consistency or inconsistency of this group of symbolic facts (colors). To denote the degree of inconsistency, we use the symbol σ . We want to compute σ .

3 Motivation

You can measure the weight of an object, its length, its volume, etc. For example, to measure the length of a door, where four workers take the measurement independently, the measures being of *2m*, *2.3m*, *3m*, and *2mts*. The must likely length is obtained by taking the average of all, which is *2.32m* in our example.

On the other hand, if the measurements are not numeric, then we have observations ("facts") of the particular event. For example, four people said:

- "Pedro's sweater is red",
- "Pedro's sweater is pink",
- "Pedro's sweater is clear",
- "Pedro's sweater is orange".

What is the color that makes more sense?, how to calculate the "average" of these facts?, how to combine the observed colors to determine which is the one that more approaches to the real color?, can we measure the degree of discrepancy among each one of these facts and the most likely real color?.

To find this "average", we place the reported colors in a hierarchy of colors. In this work we present a methodology that will allow to find the average of *n* qualitative variables.

These logic types except the diffuse logic have only two truth value, true and false, with no shades or gradations of truthfulness or falsehood. But the real world is more complicated. There are events that are not completely true or totally false, such as "the sky is blue" or "the weather is hot". Fuzzy logic solves this and provides degrees of veracity, by requiring a membership function whose range of values is [0,1] .

The Theory of Dempster-Shafer takes subjective probabilities for the observers. That is, for it people have different degrees of trust (some lie more than other).

The figure 1 shows the development of how to find the fact more commendable of a set of facts over a particular event. The observers inform of facts from a particular situation, after these facts are represented in a hierarchy and we calculate the inconsistency with the Model to measure the inconsistency (MMI).

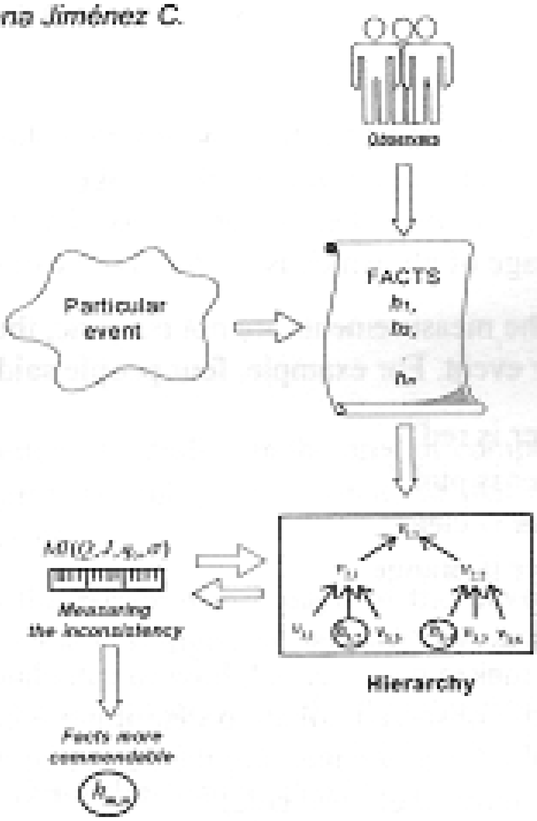


Figure 1. *Scheme to compute the inconsistency of a particular event.*

4 Previous works

Among the theories that have been dedicate to the study of the inconsistency in information, is the theory of Dempster-Shafter [10], a mathematical theory of the evidence that was introduced in the 70's and developed by Glenn Shafer and later extended by Arthur Dempster based on belief functions and commendable reasoning, which is used to combine pieces separated from information (evidences) to calculate the probability of an event.

The Theory of Dempster-Shafer is based on obtaining degrees of beliefs for a question from subjective probabilities, and combining such belief degrees when they are based on independent elements of evidences. In summary, to obtain the degree of belief, for a question (did a leaf fall in the car?) it assesses the probabilities of another question (is the testimony reliable?). The rule of Dempster begins with the supposition that the question for which it has probabilities is independent with regard to trials of subjective probabilities but this independence is only a priori; this disappears when the conflict is discerned among the different evidence elements. Contrary to Dempster-Shafer, in our work the **observers** that report on the facts have the same credibility (all say the truth) and if their facts (assertions) differ, it is due to errors or imprecisions in the observations, and not to a desire or impulse to lie. For instance, an observer saw Pedro at sunset time, so he reports "his sweater is orange", while other observer could only ascertain that "his sweater has a clear color" because the light was him.

To determine the fact with smaller inconsistency, in this work we use the hierarchies and the *Confusion* function [5], [9]. This function evaluates the similarity qualitative value *s* with regard to another *r*, both being represented in a hierarchy. For example, what is the confusion of using *dog* instead of *German Shepherd*?. We now give an example and the equations for determine the value of the function of *Confusion* (see figure 2).

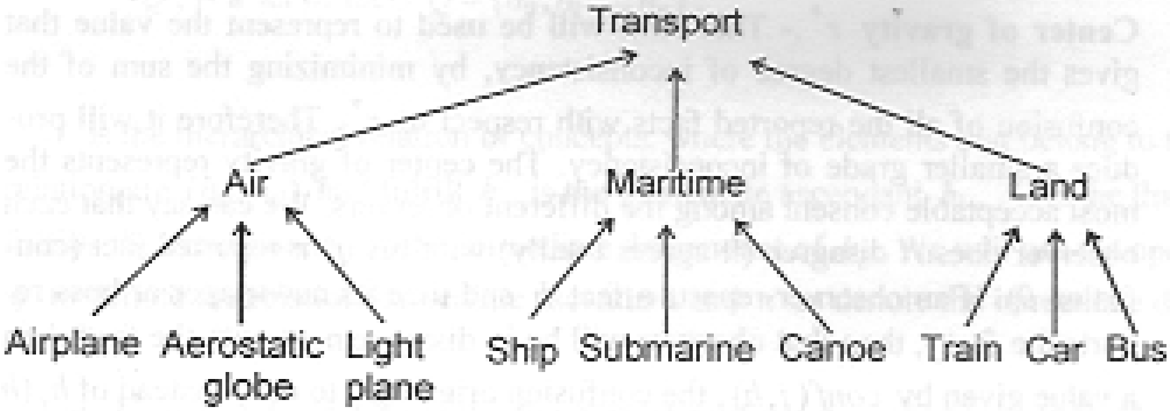


Figure 2. *Hierarchy of several types of transports for travel by air, water and land.*

The *confusion* of using *r* instead of *s*, for a hierarchy *H* (in this case the hierarchy of transports) the calculus is:

If $r, s \in H$, then the *confusion* in using *r* instead of *s*, written $conf(r,s)$, is:

- $conf(r,r) = conf(r,s) = 0$, when *s* is any ascendant of *r*.
- $conf(r,s) = 1 + conf(r, father\_of(s))$

To determine the confusion between *Transport* and *Air* is $conf(Transport, Air) = 1 + conf(Air, father\_of(Air)) = 1 + 0 = 1$. In this case, the confusion is *1*, because we are using *Transport* instead of *Aereo*. Due to the location in that, it is in the hierarchy and the rules of *confusion*, we travel the tree, where the upward levels the value is *0* and for each descent it will be *1*. Now then, if we obtain $conf(Air, Transport) = 0$, due to *Air* is a *Transport*.

Exemplifying the way to use hierarchies and *confusion* we will give a better vision of what is being carried out in this work, since it is a fundamental part of this.

5 Development

The Model to measure the inconsistency (MMI), finds the inconsistency of a set of facts. A value is calculated, which is interpreted as the degree of inconsistency, if it is close to zero means little inconsistency.

Definitions

- **Fact** (atomic fact): It is a measurement (numeric value) or an observation (symbolic) of an aspect (characteristic or property) of the reality. For example: (Juan, height, 1.77m), (Juan, hair's color, black).

- **Observer** (reporter, informant). - A person reporting one or more facts coming from particular event.

- **Center of gravity** r^* .- This term will be used to represent the value that gives the smallest degree of inconsistency, by minimizing the sum of the confusion of all the reported facts with respect to r^* . Therefore it will produce a smaller grade of inconsistency. The center of gravity represents the most acceptable consent among the different observers. We can say that each observer doesn't disagree (it agrees totally) with his own reported fact (confusion *0*). If an observer reports a fact h and then its newspaper or boss reports the fact *j*, then that observer will be in disagreement with the fact j in a value given by $conf(j,h)$, the confusion originated to use *j* instead of h (h was the reported by the observer). r^* is the fact j that minimizes the joint dissatisfaction or disagreement among the observers, or in fact, among the facts reported by the observers. r^* is the value *j* that minimizes $\sum_{i=1}^{n}(j,h_i)$, when using each reported value h_i instead of the most likely value r^* . n represents the reported observations.

- sigma σ .- It is the average of the additions of confusions. σ gives us idea of the average of dissatisfaction or disagreement that the observers have whose facts have you "summarized" reporting a single value r^* instead of $\{h_1, h_2, ..., h_n\}$. These observers reported h_i that differ something from the most likely value r^* . Each observer i has a certain dissatisfaction expressed by $conf(r^*, h_i)$. The average of those dissatisfactions is σ . n is the number of observations made from the event.

$$\sigma = \frac{\sum_{i=1}^{n}(r^*, h_i)}{n}$$

- Confusion. It has been defined in page 5 [5].

5.1 The Model to measure the inconsistency (MMI)

Let a model defined by a fourthtuple

$$(Q, J, q_0, \sigma),$$

where

Q, is a set of facts $Q = \{h_0, h_1, ..., h_n\}$.

J is the hierarchical relation of concepts, where the elements that belong to the relation are (h_a, h_b) and fulfill h_a is the immediate ascendant h_b. $J^{'}$ is the that fulfills the condition h_a is the immediate descendant of h_b. We will use the operator ϕ to denote the relation immediate ascendant and π to denote the immediate descendant.

Let $J^T = J \cup J^{'}$, it defines the function of confusion $conf : J^T \rightarrow \{0,1\}$ like:

$$conf(h_a, h_b) = \begin{Bmatrix} 0\ si\ h_a\ \phi\ h_b \\ 1\ si\ h_a\ \pi\ h_b \end{Bmatrix} \qquad (1)$$

The function $asc : Q \rightarrow \{h_a \mid (h_a, h_b) \in J\}$ is defined like:

$$asc(h_a) = \begin{Bmatrix} h_b\ si\ h_a\ \phi\ h_b \\ \phi\ si\ h_a\ \not\phi\ h_b \end{Bmatrix} \qquad (2)$$

Let $conf_A : Q \times Q \rightarrow N$ the function of confusion[1] for anyone elements that belong to $Q \times Q$ is defined like:

$$conf_A(h_a, h_z) = \begin{Bmatrix} 0\ si\ h_a = h_z \\ 0\ si\ h_z = \phi \\ 0\ si\ h_z\ \pi\ h_y\ \pi\ ...\ \pi\ h_b, h_a \\ 1 + conf_A(h_a, asc(h_z)) \end{Bmatrix} \qquad (3)$$

Let $Q' \subset Q$ the set contains the facts of interest and Q'_p the set contains the facts of interest with more than one observation. That's to say $Q'_p = \{(h, p) \mid h \in Q', p\ is\ the\ number\ of\ observations\ over\ h\ (fact)\}$:

[1] $conf_A$ is analogous to the function $conf$ that is presented in the articles with references [6], [7].

$$conf_A^p((h_a,p_a),(h_z,p_z))=\left\{\begin{array}{c} 0\,si\;h_a=h_z \\ 0\,si\;h_z=\phi \\ 0\,si\;h_z\;\pi\;h_y\;\pi\;...\;\pi\;h_b\;\pi\;h_a \\ 1+conf_A(h_a,asc(h_z))[p_z] \end{array}\right\} \qquad (4)$$

ϕ is the empty set,

p_z represents the weight of h_z ,

q_0 represents the highest node in the hierarchy and the beginning of this,

r^* defines the hierarchical value (fact) that minimizes the addition of the function of confusion:

$$\min\sum_{i=1}^{n} conf(r^*,h_i) \qquad (5)$$

σ is the value that represents the inconsistency. If $\sigma=0$ then the inconsistency does not exist in the facts h_i, that they are contained in r^* , and σ is calculated like:

$$\sigma=\frac{\min(\sum_{i=1}^{n} conf(r^*,h_i))}{n} \qquad (6)$$

σ in equation (6) can be interpreted as the confusion average that minimizes r^* .

5.2 Measuring the inconsistency of a set of facts and finding the most acceptable value

We show the way to find the degree of inconsistency of a group of facts that describe a particular event, which were provided by observers. We analyze these facts with a confusion function, which helps us to compute the center of gravity of the set of facts.

That is, the fact that generates the smallest average inconsistency or which is the most believable, could be call it also the less lying or the less erroneous.

Example

We want to determine which animal is the pet of John, when the observers reported the following facts:

John has a siamese cat
John has a siamese cat
John has a feline

John has a chihuahueño
John has a dog
John has a dog
John has a Xoloitzcuintle
John has a domestic cat
John has a eagle

Once we have the list of facts, locate them in a hierarchy (the hierarchy is designed specialized according to the knowledge of some external source, it is "general knowledge"). The figure[2] 3 shows a hierarchy J, that includes the qualitative variables that were obtained of the facts, where the observations are represented by an * (asterisk):

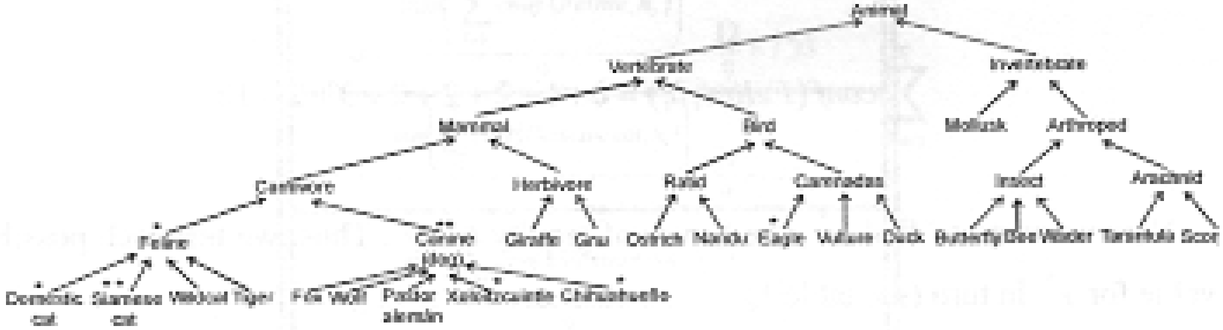


Figure 3.*Hierarchy of animals, where appear the facts from the above list.*

The set of facts is:

$Q' = \{Feline, Domestic\ cat, Siamese\ cat, Dog, Xoloitzcuintle, Eagle, Chihuahueño\}$,

but there are two observations that represented that *John has a Siamese cat* and a *dog*, where the highest node in the hierarchy is:

$$q_0 = Animal$$

To determine the possible center of gravity, we must to calculate the confusions of *Feline* with each value in the set Q' :

[2] The $conf(Feline, Siamese\ cat)$ is counted in one unit by each level that goes down in the tree from the node *Feline* to the node *Siamese cat*, the levels that ascend they don't count.

$$conf(Feline, Feline) = 0 \text{ (using Feline instead of Feline)},$$

$$conf(Feline, Domestic\ cat) = 1 \text{ (using Feline instead of Domestic cat)},$$

$$conf(Feline, Siamese\ cat) = 2 \text{ (using Feline instead of Siamese cat)},$$

$$conf(Feline, Dog) = 2 \text{ (using Feline instead of Dog)},$$

$$conf(Feline, Xoloitzcuintle) = 2 \text{ (using Feline instead of Xoloitzcuintle)},$$

$$conf(Feline, Eagle) = 3 \text{ (using Feline instead of Eagle)},$$

$$conf(Feline, Chihuahueño) = 2 \text{ (using Feline instead of Chihuahueño)}.$$

The sum of confusions of $Feline \times Q'$ is (For the others facts, the sum of confusions is obtained like *Feline*):

$$\sum_{i=1}^{9} conf(Feline, h_i) = 0 + 1 + 2 + 2 + 2 + 3 + 2 = 12$$

Now we want to find r^*, the center of gravity of Q'. Thus, we test each possible value for r^* in turn (see table 1).

$\sum_{i=1}^{9} conf(r^*, h_i)$
$\sum_{i=1}^{9} conf(Feline, h_i) = 12$
$\sum_{i=1}^{9} conf(Domestic\ \ cat, h_i) = 11$
$\sum_{i=1}^{9} conf(Siamese\ cat, h_i) = 10$
$\sum_{i=1}^{9} conf(Dog, h_i) = 12$
$\sum_{i=1}^{9} conf(Xoloitzcuintle, h_i) = 11$
$\sum_{i=1}^{9} conf(Eagle, h_i) = 29$
$\sum_{i=1}^{9} conf(Chihuahueño, h_i) = 11$

Table 1. *Candidates for gravity center of the facts Q'.*

Now, we find the value that fulfills:

$$\min(\sum_{i=1}^{9} conf(r^*, h_i)) \qquad \text{(a)}$$

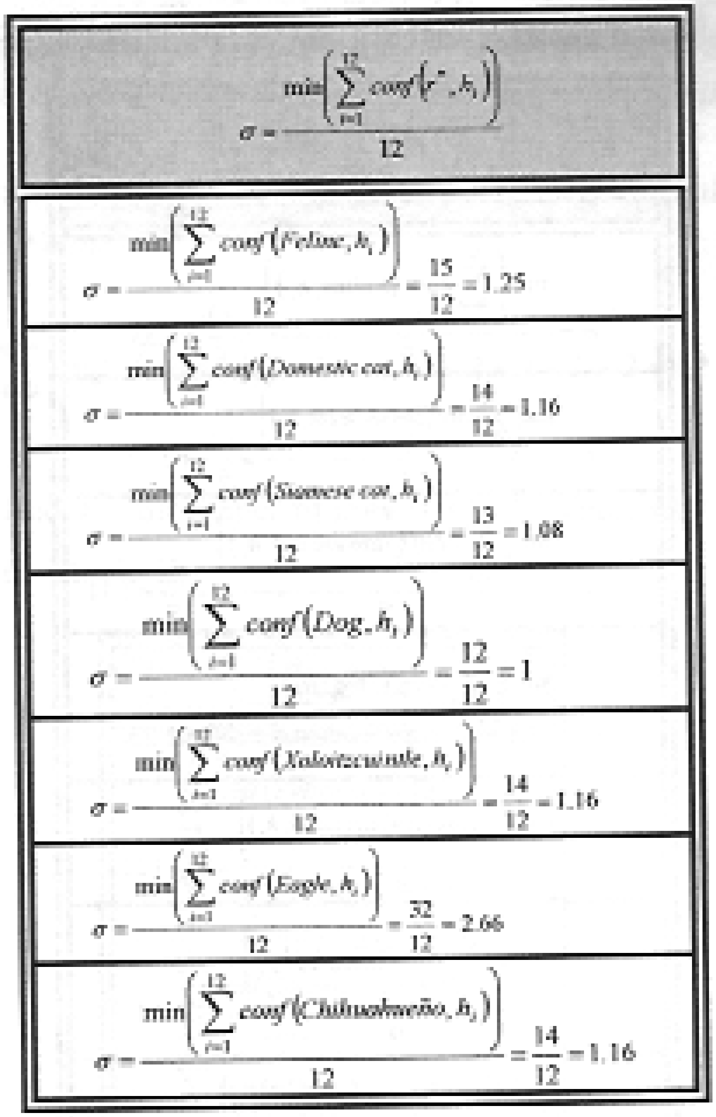
The value that fulfills equation (a) is $r^* = Siamese\ cat$ with $\sigma = 1.11$. This means that the most plausible value for the pet of John is Siamese cat, it is the value that minimizes the discomfort (measured by the confusion) of all the observers, which reported a value and "the real value published" was $r^* = Siamese\ cat$.

In the table 2, are showed the inconsistency degrees for all the elements Q'.

$\sigma = \frac{\min\left(\sum_{i=1}^{9} conf(r^*, h_i)\right)}{9}$
$\sigma = \frac{\min\left(\sum_{i=1}^{9} conf(Feline, h_i)\right)}{9} = \frac{12}{9} = 1.33$
$\sigma = \frac{\min\left(\sum_{i=1}^{9} conf(Domestic\ cat, h_i)\right)}{9} = \frac{11}{9} = 1.22$
$\sigma = \frac{\min\left(\sum_{i=1}^{9} conf(Siamese\ cat, h_i)\right)}{9} = \frac{10}{9} = 1.11$
$\sigma = \frac{\min\left(\sum_{i=1}^{9} conf(Dog, h_i)\right)}{9} = \frac{12}{9} = 1.33$
$\sigma = \frac{\min\left(\sum_{i=1}^{9} conf(Xoloitzcuintle, h_i)\right)}{9} = \frac{11}{9} = 1.22$
$\sigma = \frac{\min\left(\sum_{i=1}^{9} conf(Eagle, h_i)\right)}{9} = \frac{12}{9} = 1.33$
$\sigma = \frac{\min\left(\sum_{i=1}^{9} conf(Chihuahueño, h_i)\right)}{9} = \frac{11}{9} = 1.22$

Table 2. *Inconsistency degrees for the elements of Q'*

In our example, r^* turned out to be the most specific fact (the fact deepest inside the hierarchy the fact furthest away from the root). This is not always the case. If five observations (facts) reporting *Dog* would have made $r^* = Dog$ with $\sum_{i=1}^{12} conf(Dog, h_i) = 12$ and $\sigma = \frac{12}{12} = 1$. For the rest of facts, the degrees of inconsistency are the following:



$\sigma = \frac{\min\left(\sum_{i=1}^{12} conf\left(r^{*}, h_i\right)\right)}{12}$
$\sigma = \frac{\min\left(\sum_{i=1}^{12} conf\left(Feline, h_i\right)\right)}{12} = \frac{15}{12} = 1.25$
$\sigma = \frac{\min\left(\sum_{i=1}^{12} conf\left(Domestic\ cat, h_i\right)\right)}{12} = \frac{14}{12} = 1.16$
$\sigma = \frac{\min\left(\sum_{i=1}^{12} conf\left(Siamese\ cat, h_i\right)\right)}{12} = \frac{13}{12} = 1.08$
$\sigma = \frac{\min\left(\sum_{i=1}^{12} conf\left(Dog, h_i\right)\right)}{12} = \frac{12}{12} = 1$
$\sigma = \frac{\min\left(\sum_{i=1}^{12} conf\left(Xoloitzcuintle, h_i\right)\right)}{12} = \frac{14}{12} = 1.16$
$\sigma = \frac{\min\left(\sum_{i=1}^{12} conf\left(Eagle, h_i\right)\right)}{12} = \frac{32}{12} = 2.66$
$\sigma = \frac{\min\left(\sum_{i=1}^{12} conf\left(Chihuahueño, h_i\right)\right)}{12} = \frac{14}{12} = 1.16$

Table 3. *Inconsistency degrees for the elements of Q'.*

Conclusions

This model allows us, (1) to find the inconsistency in a set of facts; (2) to compute the degree of inconsistency of a set of facts. In (1) and (2) are carried out using hierarchies, instead of assigning subjective probabilities to the truth (reliability) of the observers, as Dempster-Schafer does or values that in some given moment they take us away from the reality of the facts.

Therefore, we can find the most commendable fact of a particular situation and a serie of inconsistency degrees. We no longer assert "these facts are inconsistent" or "these facts are consistent", as classic logic does. Now, we can say "these facts are consistent in degree x", where $x > 0$.

An obstacle can be that more complex facts are not managed, but that will be a future work.

References

1. Byrne, Emma; Hunter, Anthony. Man Bites Dog: Looking for Interesting Inconsistencies in Structured News Reports. Department of Computer Science. University College London, Grower Street. May 29, 2003
2. Carlson, Jennifer; R. Murphy, Robin: Use of Dempster-Shafer Conflict Metric to Detect Interpretation Inconsistency. Proceedings of the 21th Annual Conference on Uncertainty in Artificial Intelligence (UAI-05), AUAI Press, Arlington, Virginia. Pages 94-104. 2005
3. Easterbrook, Steve: Learning from Inconsistency. International Workshop on Software Specifications and Design. Proceedings of the 8^{th} International Workshop on Software Specification and Design. Page 136. 1996.
4. Gabbay, Dov; Hunter Anthony: Making inconsistency respectable 1: A logical framework for inconsistency in reasoning. In Fundamentals of Artificial Intelligence Research, volume 535 of Lecture Notes in Computer Science, pages 19-32. Springer, 1991
5. Guzmán-Arenas, Adolfo; Levachkine, Serguei. Relatedness of the Elements of Hierarchies Partitioned by Percentages. Center of Computing Research- National Polytechnic Institute CIC-IPN.Ronny Lempel, Shlomo Moran, *Optimizing result prefetching in web search engines with segmented indices*, ACM Transactions on Internet Technology (TOIT), Volume 4 Issue 1, February 2004
6. Guzmán, A., and Levachkine, S. (2004) Hierarchies Measuring Qualitative Variables. Lecture Notes in Computer Science LNCS 2945 (Computational Linguistics and Intelligent Text Processing), Springer-Verlag
7. Hunter, Anthony: Measuring Inconsistency in knowledge via Quasi-classical Models. Eighteenth national conference on Artificial intelligence, p.68-73, July 28-August 01, 2002, Edmonton, Alberta, Canada
8. Hunter, Anthony: Reasoning with Contradictory Information using Quasi-classical Logic. Journal of Logic and Computation, volume 10, No. 5, 677-703, 2000
9. Levachkine, Serguei; Guzm\'{a}n-Arenas, Adolfo; Polo-de Gyves, Victor: The semantics of confusion in hierarchies: Theory and practice. In Proceedings of the 13th International Conference on Conceptual Structures: common semantics for sharing knowledge (ICCS 05). Kassel, Germany, 2005
10. Shafer, Glenn. A Mathematical Theory of Evidence. Princeton, N. Y. Princeton University Press, 1976
11. Shafer, Glen. Rejoinders to Comments on "Perspectives on the Theory and Practice of Belief Functions". International Journal of Approximate Reasoning, volume 6, No. 3, 445-480, 1992