Towards a Theory of Quantity

Praveen Paritosh
Department of Computer Science, Northwestern University, Evanston, IL 60201, USA.
Tuesday, October 29, 2002

1 Introduction

Quantities are ubiquitous and an important part of our understanding about the world. In this paper, I present a theory of quantity – representations and principles for generating those representations. The notion of quantity is quite broad, and there is significant psychological literature on many of specific aspects of it. There's quite a rich literature on the psychology of perceptual quantities like brightness, loudness, etc., (the literature refers to them as *dimensions*), and not much on conceptual quantities like price of computers, GPA of students, etc. Most of the research (from the decision making community) that does study conceptual quantities employs feature vector models. On the other hand, the structured models of similarity and generalization do not handle quantities adequately.

This paper presents a theory of quantity that naturally fits with the structured and symbolic framework of knowledge representation and structure based models of similarity and generalization. The work presented here would be a key component of doing back-of-the-envelope reasoning, which involves making quick, quantitative estimates. Far from being rigorously psychological, the spirit is to provide a natural extension to known cognitive models, and try to go as far as possible on the dimension of cognitive plausibility, in the service of building a back-of-the-envelope reasoning system.

The next section gives examples of what I mean by quantities. Section 3 presents the motivations driving the work. A symbolic partitioning of the space of quantity is presented as a representation for quantities which is discussed in section 4. Section 5 presents the principle underlying these natural partitions, namely, things in the world come in *structural bundles.* The last two sections talk about how to test these ideas, and their implications.

2 Quantities

Quantities occur in knowledge about many kinds of things – cars have engine horsepower, size, mileage, price; countries have GDP, population, area; birds have wingspan, weight, surface area, and so on. The notion of quantity is usually employed to describe attributes/features that take on values that vary continuously over an ordered space. Thus, we have two types of attributes – attributes that take *nominal values* (e.g., race, sex, color), and attributes that take *quantitative values* (e.g., size, temperature, price). The latter kind of attributes will be referred to as *quantities* in this paper[1]. Things that one can do with quantities[2] that aren't possible with nominal attributes are –

[1] Whether something is a quantity or a nominal attribute depends upon the representation, as one can think of color as a nominal attribute that's not a quantity, when one thinks of it taking symbolic values like red, blue, green, etc; but if one represents it as hue and saturation, these are full-fledged quantities.
[2] This is an age-old distinction – nominal, ordinal, interval and ratio scales; based on what arithmetic operations are supported. A ratio scale is also interval and ordinal, and an interval scale is also a ordinal scale, but the converse is not true.

- Ordinal comparisons: A Honda Accord is more expensive than a Kia Rio.
- Differences: The difference in prices of Maserati Spyder and Porsche Boxter is much more than that between the Boxter and Lexus.
- Ratios/parts: A Honda Accord is twice as expensive as a Kia Rio.

A convenient and very powerful representation of quantities is using numbers (and the relevant units). A quantity is any of the examples above, and has at least three parts -- {value, unit, attribute[3]}, for example my stipend might be {1102, USD/month, stipend}.

The way in which we acquire knowledge about the world leads me to make the following distinction between two classes of quantities –

- Perceptual: e.g., brightness, loudness, size of things that we see, etc. We have a direct sensory measure of these attributes.
- Conceptual or Abstract: e.g., prices of things, mileage of cars, gross domestic products of countries, the size of countries, the GPA of a student, the clock speed of a computer. The distinction is that there is no direct sensory perception of these attributes.

Sometimes, though, it might not be clear whether a quantity's underlying model in the person's head is perceptual or conceptual, though, as one might *perceptualize*[4] even the most abstract quantities – e.g. the size of a country as the size on the map, or the clock speed to the perceptual experience of speed of the experience of working on that computer. In this paper I will focus on the conceptual dimensions. I conjecture that rich structured representations and higher level cognitive processes play a bigger part in our understanding of conceptual dimensions than perceptual dimensions.

3 Motivation

In most symbolic knowledge representation frameworks, quantities are not handled adequately, if at all. Representing them as numbers does not go far in being useful, for example, our models of surface similarity/ shallow retrieval (MAC), similarity (SME) and generalization (SEQL) do not care much about quantities. The way quantity might be implicated in these processes –

- Surface similarity: Just as `Red` the symbol occurring in the probe might remind me of other red objects, the a bird with wing-surface-area of 0.272 sq.m. (that is the Great black-bucked gull, a large bird) should remind me of other large birds. One way to make that happen might be to abstract the numeric representation of wing-surface-area to a symbol, say, `Large`, then it will show up in content vectors, and contribute to the dot product.
- Similarity: A model of similarity that is sensitive to quantities will explain how –
    - Quantity values can make things that have similar amount of structural overlap more or less similar. There are two questions here – how to compute similarity along a single dimension, and how to combine the similarity along different dimensions in computing overall similarity of two cases. Feature vector models answer the former by computing a difference, which does not explain how distances that are close, but across a landmark,

[3] It is for example, important to distinguish between stipend, and per capita income of a country, even though the units are the same.
[4] This is related to the act of visualizing or our processes of imagery for even abstract notions.

are perceived to be more different than distances within the same qualitative region. For combining differences along more than one dimension, the feature vector models posit weights, and some weighted distance metric, but do not provide a principled way to find these weights.
    - To make estimate of an unknown quantity based on a similar known case.
- Generalization: Key part of learning a new domain is acquiring a sense of quantity for different quantities. E.g., from a trip to the zoo, a kid probably has learnt something about sizes of animals, their shelters, etc. As we learn to cook, we get a sense of amounts of ingredients, cooking times, etc.

A better understanding of quantities will help me in figuring out how to make estimates, which is a key aspect of back of the envelope reasoning. The key questions are –

- What is a cognitively-plausible representation of quantity that will let us extend our models (MAC/FAC/SEQL) to be able to do the things pointed above?
- How do we build that representation from our experiences? Or, how do we learn the sense of a quantity?

4 Representing Quantity

Any quantity that can take on continuous values in a range can has an infinite possibilities of what it can be. Clearly, our representations are not that fine-grained, and we tend to clump together parts of the space of quantity which we give names like `Large, Medium, Small,` etc; or recognize special points on the space like `Poverty Line, Freezing point`. These symbols, and the ordinal relationships between them constrain and define a representation for that quantity. This is exactly the same as the *quantity space* representation from QP theory. QP theory showed us that a lot of powerful reasoning could be done with such a representation. The symbolic and relational nature of this representation automatically makes it much more useful in our representational framework, and I conjecture that this will make our models of similarity and generalization be more aware of quantities.

The proposed representation is a set of symbols on the quantity space that partition the space of quantity – sometimes the symbols correspond to points e.g., `Poverty Line, Freezing Point`, and sometimes to intervals e.g., `Upper Class, Large`. The symbols may be partially or totally ordered. Given this framework for representing quantities, we still have to figure out what and where are these special points[5] on the space of quantity.

One should only make the *necessary* and *relevant* qualitative distinctions, QP theory advises us. In the domain of processes, QP theory provides the intuition for these distinctions: *where things change*, i.e., different processes and/or model fragments get de/activated, e.g., `Freezing Point` and `Boiling Point` of a liquid. But the *quantity space* representation seems to be more general than just processes – e.g., our notion of cheap, medium-range and expensive computers, or division of income groups into poverty line, lower class, middle class and upper class, etc. Is there a more general

[5] These have been called *landmarks* and *benchmarks* in the literature. QP theory makes a very important distinction between landmarks and limit points, which amounts to *landmarks* being concrete instantiations of *limit points*. Right now, lets be non-committal to that, and call these the *partitions* of the quantity space.

notion that will provide us a principled way of finding the necessary and relevant distinctions? Of course, one can always partition the quantity space arbitrarily – so, one could have an ad hoc rule that said that we'll always divide the space between the minimum and maximum into three parts – high, medium and low[6]. That would mean that I am suggesting that there are some partitions that are more *natural* than others. In the next section I discuss about the nature of these partitions, and give some examples that lead to the intuition behind the principle that might be underlying these *natural partitions* of the space of quantity.

Of all possible ways to partition the quantity space, some are more natural. These are those that correspond to the *necessary* and *relevant* qualitative distinctions. If I present a principle for finding natural partitions, it is my responsibility to show why they are natural. However, it is not easy to provide metrics for why they are more natural than others; below I will present some features of the natural partitions –

- Just the right level of granularity for the kind of comparisons and reasoning one does with the knowledge. E.g., [`Freezing Point, Boiling Point`] might be fine for reasoning about physical behavior, but if one is talking about shower water, then [`Cold, Body Temperature, Warm, Scalding Hot`] might be more appropriate.
- Predictive of other properties of the system. E.g., [`Poverty Line, Lower Class, Middle Class, Upper Class`].

The origins of these partitions of the quantity space are very diverse, e.g. –

- Personal experiences e.g., one's notion of `[Bland, Mild, Spicy].`
- (Social) conventions e.g., [`Poverty Line, Lower Class, Middle Class, Upper Class`], or `[Short, Regular, Tall]` when talking of people's heights (this seems to be part social, part personal experiences).
- Physics of the phenomenon - [`Freezing Point, Boiling Point`].

5 Structural Bundles

Consider any real world object[7], most of the attributes that describe it are constrained, in the sense that they can not take on any arbitrary values. There seems to be two major classes of constraints on the values that a quantity can have –

- *Distributional Constraints*: Most quantities have a range (a minimum and a maximum) and a distribution that determines how often a specific value shows up. E.g., the height of adult men might be between 4 and 10 ft, with most being around 5-6.5ft.
- *Relational Constraints*: Besides the above, a quantity is also constrained by what values *other* quantities in the system take, its relationship with those other quantities, the causal theories of the domain; in general, the underlying structure of representation. Lets look at some examples –

[6] For example, the Fuzzy logic community does something in the same spirit.

[7] Comic books, mythology, and fantasy, for example, has the freedom to relax this constraint – a character can be arbitrarily strong, large, small or be able to fly, even though the physical design of the character might not be able to support it.

- This generally holds for all internal combustion engines (bikes, cars, planes, etc.). As the engine mass increases, the Brake Horse Power (BHP), Bore (diameter), Displacement (volume) increases; and the RPM decreases.
- As the length of an animal increases, the length of the vocal chords increases. Thus, the larger the animal, the deeper its voice. Of course, that has implication for the entire sound producing/receiving apparatus of the animal, the distance its sound travels, etc.
- The surface area of an animal determines the heat loss/exchange of gases/assimilation of food that it is capable of. Therefore, all spherical organisms are smaller than 1 mm in diameter, as the sphere is the shape with minimum surface area per unit volume. For animals that swim, it determines the drag, and for birds, the lift that it can generate. So, a change in surface area has repercussions on many aspects of the animal.

The *relational constraints* on quantities reflects a fundamental fact about the way things are in the world. Things in the world come in *packages* or *bundles.* For example, a "muscle car" has a powerful engine, is expensive, is designed for style and fun rather than practical driving. In literature, a similar notion is expressed by *attribute co-variation* or *feature correlation.* But there's much more than that – these are not merely bundles of correlated attributes, but are *structural bundles*[8]. The entities, and quantities associated with them, tied by relations and higher order relations constraining them, give rise to the structure therein. Processes (as in QP theory), are a special case of these *structural bundles* for the class of dynamical physical systems. The necessary and relevant qualitative distinctions correspond to discontinuities in the underlying reality as captured by the structure in the representation[9]. Different quantities vary in the degree to which they are relationally constrained – let us call a quantity that participate in more deeply nested parts of representation as more *systematic* than another which participates in less deeply nested parts of representation.

The above observation is the key to the principle underlying the natural partitions of the quantity space. Let's take an example – people's income. We have a set of exemplars of people from different classes which consist of structured representations of the social and financial aspects of their life, and income is an aligned dimension across all the exemplars. Imagine that we have ordered the exemplars by income and as you're walking from the minimum income exemplar to the maximum income exemplar. My bet is that as you make transition from lower to middle, and middle to upper, there will be a discontinuity in the structural similarity of the exemplars. Not only are the incomes of two exemplars in the lower income group similar, but they share much more structural similarities then one taken from the lower class and the other from middle or upper class.

[8] The idea of bundles isn't new, Rosch has talked about it, and maybe it's much older. But since their bundles were of flat attributes, from a feature vector, the explanatory power of those bundles, in my opinion, is far less than structural bundles.

[9] What if there are none? For example, imagine a set of exemplars whose structured representations are identical, and the only difference is in the value of the quantities. I can only imagine this happening with poorly represented examples (as one is just beginning to learn about a domain, for example, one might think that all automobiles are exactly the same but for their sizes and colors) or artificial examples. This where the *distributional constraints* might provide the partitioning. More about this in the next section.

Let's call the partitions provided by the idea in the above example as *relational limit points*. Starting from a set of exemplars[10], the relational limit points are the ordered transitions on the space of alignable quantities that correspond to structurally distinct clusters. The claim here is that these relational limit points are a major part of the natural partitions on most quantity space.

6 Hypothesis Testing

To support, test and refine the ideas presented above, currently I am trying to implement them, to generate the *natural partitions*. The CYC knowledge base contains a reasonable amount of information about countries (taken from the CIA World Factbook, there are an average of 250 assertions for 200 countries). There are about thirty quantities whose numeric values are known for most countries (`area, population, birth rate, total labor force, gross domestic product, number of airports, etc.`). Currently there isn't much knowledge about the relationships between these parameters, but I will be adding that. There are many other facts about different countries – like the organizations that it's a member of, different political events in which it was a part, etc. that are also in the knowledge base. The first step is to build a case library of a large number of countries that has lots of quantities, and reasonably rich in structure. Once we have that, the goal is to see if we can use the above ideas to find the *natural partitions* for the quantities, and see how that gives us better performance with MAC/FAC, and coming up with estimates for missing quantities.

The first problem is to bootstrap the process. To find the interesting relational limit points, we cannot ignore the numbers altogether, and thus we need to a do a rough, first-cut symbolic partitioning for the quantities. In the previous section, we mentioned distributional and relational constraints. This is where the distributional constraints play a role. Purely by looking at the distribution of a single quantity, one can divide it into ranges (e.g., low, medium, high population). There are three major types of distributions that capture most of the real world parameters – Zipf/power law (many small countries, and a very few, but quite large countries), normal (most people are medium-height, few short, and few tall), multimodal. At the end of the paper, there is a rough sketch of an experiment that tries to test how people do this. Right now, I am trying to find a simple set of heuristics that provides a first cut partitioning into the distributional landmarks.

The simplest strategy for finding the relational limit points might be to find the structural clusters, and see if they partition the space of the quantity in an ordered manner. Or, one might present the exemplars in the order sorted by the value of a quantity, and use the resulting generalizations to partition the quantity's space. Along with that, storing the distribution of quantity values from a set of exemplars that have been assimilated in a generalization might help in giving a finer sense of quantity, and mapping back the symbol to a number.

[10] The notion of relational limit points seems to be most sensible when we start with a set of exemplars from the same super-ordinate category (in Rosch's sense of the word).

7 Implications

The distributional landmarks and relational limit points provide a symbolic and relational representation of quantity that has the potential to be quite useful in our framework, making both steps of MAC/FAC more sensitive to quantities. It will provide for a principled way to partition the space of a quantity, tells us which quantities are more predictive (those that are more *systematic*) than others of structural properties, and a simple principle for combining differences along different quantities. There are a lot of things that are not explained, some of them are –

- How do we make comparisons and quantitative judgments that cut across different categories? For example, cheap, medium-range, and pricey gifts, will consist of things without a lot of structural similarity.
- Where exactly are the symbols that make up the quantity space and the ordinal relations between them stored? If they are in the generalizations, then how do they help MAC/FAC for the exemplars? On the other hand, if they are with the exemplars, one will have to update them all in batch, as the relational limit points are dynamic, since they might change with the set of exemplars that one has seen[11].
- What about quantities that are least systematic? We wont be able to find much relational limit points for those, so the prediction here will be that most of the understanding of such quantities will be based on the distributional landmarks.
- How does the mass/count distinction effect the above ideas?

Clearly, a full understanding of quantities is a major enterprise, but hopefully the ideas here will form a part of it.

[11] The contextual determination of the semantic congruity effect (Cech and Shoben, 1985) seems to be evidence for the fact that we do not have a stable partitioning of quantities into set categories in the long-term memory.

Appendix 1: Distributional Abstraction of Landmarks[12]

There are N (say 30) circles of varying sizes. They are all alike, but for size. The subject is presented with this pile, and asked to divide it into smaller piles/groups.

A. They are free to chose the number of groups and size of each group (the number of circles in a group).
B. They are given the number of groups (e.g., 3, which could be thought of as the small, medium or large groups), but are free to chose the size of groups.

Now, one can do this, and vary the underlying distribution of the radius (or area) of the circles -- they could come from uniform, gaussian, exponential, power law, distributions. For example, if the sizes are linearly increasing between a MIN and MAX, then an equal split into N/3 circles per group might be meaningful; but if they come from a gaussian, which reflects in the fact that there are a larger number of medium-sized circles than there are of small or large. One can also try bi-modal distributions and so on. Statistically, the distribution provides some clue in partitioning. By looking at how people do in experiment A and B above, one can see how well people intuit the underlying distributions. Other key thing is a lot of real world dimensions are power-law like, e.g., there are many people with modest income, and few with arbitrarily high incomes. So, what we have done here is given people just one dimension, and we are seeing what kinds of partitioning they do.

[12] The experiment is described to give a better feel of the problem here, I'd like to gather evidence from literature for intuitions regarding how it will turn out to be, but no luck till now. I did ask Linda Smith, and she was unaware of any such experiment; and Dedre had earlier asked Satoru Suzuki, who had the same answer.