Problems with the Calculation of Novelty Metrics

David C. Brown
Worcester Polytechnic Institute, Worcester, MA, USA

Shah, Vargas-Hernandez & Smith [2003] (SVS) have proposed a widely adopted and very influential set of metrics, including one for novelty. They attempted to make them specifically appropriate for Engineering Design, while keeping principles from Cognitive Psychology in mind. In this position paper we view the SVS proposal for possible use in Computational Design Creativity systems. Unfortunately we conclude that, while it appears to be useful for human assessment of design novelty it poses significant problems for computer use.

1. Introduction

In the many lists of aspects that creativity researchers associate with a creative product, "novelty" is king. While not everyone agrees about what novelty means precisely, it is the most obvious feature of a creative product. The number of aspects written about varies from a handful to a giddy thirty indicators of creativity [Besemer 2006], [Cropley & Kaufman 2012].

However, very little attention has been paid to how these aspects might be used in Computational Design Creativity (CDC). See Maher & Fisher [2012], Brown [2012] [2013] [2014], and their references, for a discussion of this.

Most references discuss the aspects assuming that *humans* (often experts), not computers, will be making a judgment, using either a qualitative or quantitative score. For example, one technique, proposed by Charyton et al. [2008], has human evaluators of product novelty select a keyword (e.g., "Commonplace") from a given list to try to summarize their feelings about the product's novelty. Each keyword in the list has a numerical "originality" value associated with it, and that value is then used as the value for the novelty metric (e.g., higher values represent more novelty).

Design Computing and Cognition DCC'14. J.S. Gero (ed),
pp. xx-yy.

Clearly it involves an action that a computer can't do. Additionally, there are two biased selections (the choices of the words for the list and the choices of the actual values) as well as two potentially inaccurate heuristic matches (between the product and the word, as well as between each word and its value). While it is deeply flawed even for human use, it may still be acceptable as long as it is consistently applied and its results match human novelty judgments closely enough to be acceptable. Note that if this sort of method can be adapted for computer use, the same flaws will be present, but the same acceptability conditions should apply.

In Engineering, calculated metrics are preferred over keywords, giving the (usually false) impression of precision, and allowing the possibility of computing an average novelty for a set of designs: often a set of competing conceptual designs for the same problem.

Having a metric that can be consistently applied allows comparisons between designers, and allows experiments to investigate the impact of different stimuli or procedures on design creativity.

Shah, Vargas-Hernandez & Smith [2003] (SVS) proposed a widely adopted and very influential set of metrics, including one for novelty. They attempted to make them specifically appropriate for Engineering Design, while keeping principles from Cognitive Psychology in mind. In this paper we view the proposed SVS novelty metrics, and subsequent variations, for possible use in CDC systems.

2. The Base Methods

The SVS proposals have been very influential, with chain of both users and modifiers (i.e., those who have proposed modified metrics). SVS state that "novelty is a measure of how *unusual* or *unexpected* an idea is compared to other ideas". Note that to be "unexpected" depends on expectations, that expectations can be frequency based, and that surprise can be based on violated expectations, so SVS's view allows a connection between novelty and surprise. However, expectations can be formed other ways, and novel items may not be surprising [Brown 2012] [Maher & Fisher 2012]. It is much better to separate novelty and surprise. In addition, "unusual" can also be seen as "different". Difference detection poses its own problems.

Shah uses four levels for "variety" estimation: physical principles, working principles, embodiment & detail. However, these four levels are not used for novelty: the design stages "conceptual" or "embodiment" are used instead. Hence, the paper is biased towards conceptual designs, not detailed, completed designs.

2.1. Novelty Method 1

SVS introduce two ways to calculate a novelty metric. The first, estimated by experts, is based on all the existing ideas for the given problem. As a consequence it is referred to as the *a priori* method. While it is hard to find "all" the existing ideas, in order to use them "for comparison" they must also be relevant ideas: even harder to determine. This is the 'comparison set'.

Using the comparison set, the expert determines a set of "key" attributes (i.e., "functions or characteristics") at the conceptual and/or the embodiment level. An example attribute is "method of motion", with possible values such as "wheels" or "fly". Given the comparison set, and the key attributes, the goal is to determine what is unusual (i.e., novel).

For each value found for each attribute, a novelty score (e.g., 3, 7, 10) is assigned. More frequent values get lower scores, while unusual values get higher scores: i.e., less frequent = more unusual = more novelty.

This preset range of novelty scores are a proxy for frequency of occurrence of attribute values in the comparison set. This suggests the possibility of a more nuanced scoring system linked to frequency.

Note that finding the novelty score for a value may involve finding the closest match: e.g., is "hover" closer to the possible values "fly", "jump" or "slide"? It isn't clear how that is to be done: even less so when considers doing it computationally.

To add more subtlety to the method, each attribute is given an "importance" weight f_i. This weight makes the most sense for functions.

In addition, each stage (i.e., conceptual and embodiment) is given an "importance" weight p_k. As more commitment to the solution is typically made at the earlier stages, it probably makes sense for conceptual novelty to be weighted more highly. Sarkar & Chakrabarti [2011] make a similar argument.

2.2. Issues with Novelty Method 1

Some issues related to computational use have already been raised above. Issues range from general problems to potential problems for a CDC system. The issues include:

- How to determine the comparison set.
- How to determine which attributes are "key". Note that SVS require all of the relevant design ideas to share all of these key attributes: i.e., those are "required" by the design problem.
- How to find the functions in the design representation.
- How to determine the values that may possibly be novel.

- How to determine how many novelty score values there should be.
- How to judge which novelty score is appropriate for each value of an attribute.
- How to manage any similarity judgments between values.
- How to determine appropriate use of "other" to account for infrequent attribute values (i.e., how infrequent?).
- How to determine the impact of this being a context free method? e.g., two values might be very common independently, but very uncommon together (e.g., slide *and* wheels).
- How to determine the weights for the attributes (or functions).
- How to determine the weights for the stages.
- How sensitive the novelty evaluation scheme is to the values for these weights.
- How sensitive the novelty evaluation scheme is to the novelty score values.

2.3. Novelty Method 2

The second novelty metric may have more potential for automated calculation. It is based on the set of ideas (a comparison set) generated by participants, typically in a design experiment. As a consequence it is referred to as the *a posteriori* method.

Once again it starts by identifying the key attributes of the ideas: typically functions. Then all the values are found that have been produced for those attributes. Next count how many instances of each value there is. This is done for both the conceptual and embodiment stages.

This produces m functions and n stages (n = 1 or 2), where f_i weights represent function importance, and p_k weights represent the importance of each stage. The method aims to produce a weighted score across all functions and stages for a single design.

The novelty score for a value for a single function/attribute at a single stage is based on ((T – C) / T) where T is total number of values produced for the function in that stage, and C is the number of times the value from this design (e.g., fly) was used in the comparison set. This means that if the value is "rare" C will be small and ((T – C) / T) will be closer to 1; i.e., the score is based on *rarity*, relative to all the values produced by the participants. As a frequency based measure, it seems to be a measure of how capable the participants are of producing unusual values for attributes: i.e., it is a measure that is *relative* to and sensitive to the group.

Note that deciding that a value belongs in the C count depends on the fact that the values in T are probably already normalized in some way

(e.g., "almost square" is considered to be "square") based on difference/similarity.

2.4. Issues with Novelty Method 2

Some issues have already been raised above. Several that apply to method 1 also apply here. The issues include:

- The need to define what conceptual and embodiment design descriptions look like.
- The need to have those descriptions for all the designs actually being evaluated.
- The method is restricted to the set of designs by participants.
- The method requires all designs to solve the same problem.
- The novelty is relative to the comparison set.
- The novelty is limited to being relative to the designers involved.
- All weights need to be determined.
- Counting T and C for a large comparison set will be difficult.
- It is unclear why is metric isn't based on more levels of design description (e.g., FBS) or design stages.

2.5. Refinements

In what follows below, we will briefly examine work done that is based on, attempts to refine, or comments on, SVS. Concentrating on the metrics for novelty we try to ask about issues such as whether the method is precise enough to be done computationally; whether the method scales up; what stance it takes regarding the measurement of novelty; and which of the issues above their modifications affect.

3. Nelson et al.

Nelson et al. [2009] propose refined metrics for measuring ideation effectiveness, focusing on the "variety" metric. They note that as novelty can be calculated for a single design then average novelty can be calculated for a set of designs, or for each function in a set of designs. A problem is that average novelty across *any* set doesn't make sense as the items in the set need to be related in some way. Their critique of the novelty metric is that it is "essentially a measure of whether the exploration occurred in areas of the design space that are well-travelled or little-travelled."

4. Srivathsavai et al.

Srivathsavai et al. [2010] check novelty at the whole *concept* level and at the *feature* level, where features are considered "basic" or "additional". Their first approach is based on SVS's second method.

At the 'concept' level, novelty is once again based on ((T – C) / T), with T = total number of designs, and C = number of designs of the same type. Note that "same" is a judgment that depends on having a standard set of types. In the alarm clock example they provide the type comes from features "Music", "Shape/layout" and possibly "Display type". These types are also a judgment.

At the 'feature' level, basic features are those that are "essential" and related to main function of product: e.g., the "mode of alarm" for an Alarm Clock. Features are determined by examination of currently available products. SVS's 2nd method is used with the features: once with just basic features, once with all features. For each feature, "commonly found" values were identified as "standard" values. This assumes that every design shares the same set of features. Note that a label of "standard" might actually include different values, which adds error. There's also the issue of what "common" means.

At the 'feature' level novelty is based on a sum of the novelty of each feature, also using novelty as ((T – C) / T) for each feature. Srivathsavai et al. found that feature level novelty is more repeatable between evaluators.

The second approach by these authors, an Originality Metric, is based on Charyton et al. [2008]. An 11 point scale is used to reflect increasing "originality" with scores corresponding to originality terms (e.g., Dull). A human judge selects an appropriate term, and retrieves the score. The terms are: Dull; Commonplace; Somewhat Interesting; Interesting; Very Interesting; Unique and Different; Insightful; Exceptional; Valuable to the Field; Innovative; Genius.

The first problem with this scale is that it depends on the meanings understood by the person judging, with probably inconsistencies in interpretation between evaluators. The second problem is that it isn't clear how much difference there is between some terms, such as Dull/Commonplace and Insightful/Exceptional, for example. The third problem is that people tend to have trouble deciding between anything more than a handful of values (e.g., 5-7). The fourth, and probably the most important problem, is that these terms are actually from *different* scales. These scales include the degree of affecting/engaging the evaluator's attention, the frequency of occurrence of this solution, the degree of expertise of the designer, and the value/utility of the design.

Apparently recognizing (some of) the problems with the 11 point scale, Srivathsavai et al. also try a 4 point and a 3 point scale. Unfortunately, despite improved inter-rater reliability, these scales share some of the problems outlined above. It should be clear that the originality scales are hard for a human, and impossible for a computer to use. Unless the underlying scales can be made computational, then there is little chance of CDC system use.

Some of the issues include:

- Why distinguish between basic and additional features?
- How many products need to be inspected in order to determine what features to use?
- How was a feature value determined to be "common"? i.e., what is the frequency/percentage cutoff?
- Why weren't basic features given more weight?
- They acknowledge that the novelty metric does not compare the ideas to past ideas or current products.
- They acknowledge that existing metrics focus on functions (and technical features) while non-functional features (e.g., user interaction [Saunders et al. 2009]) contribute to evaluation (also indicated by Besemer [2006]).

5. Peeters et al.

Peeters et al. [2010] refine of SVS's novelty metric by splitting the "conceptual" stage into "physical" and "working" principles, while keeping the "embodiment" stage. This shadows the Pahl & Beitz [1996] notion of "original", "adaptive" and "variant" designs, and addresses one of the criticisms of SVS. They make a physical-, working-, embodiment-based hierarchy, similar to that used in the SVS Variety metric. That hierarchy summarizes all the designs in the current set (whatever that may be), including the design in question.

A rarity score for each level counts how many examples there are: e.g., eight of one physical principle and two of another. Rarity scores for levels are weighted, combined, and normalized, to give a design's novelty score. However, level weights are needed and judgments must be made about the categories used in each hierarchy level.

Novelty scores might be high for one level, but not for another. This could indicate where the resulting novelty is coming from. This might be used to obtain better feedback about which CDC novelty introduction mechanism is being more successful.

6. Verhaegen et al.

Verhaegen et al. [2012] subdivide Novelty into "originality" and "paradigm relatedness" [Dean et al. 2006]. Originality is 'rare + ingenious, imaginative or surprising'. Rarity is fairly easy to measure (e.g., by SVS) and usually corresponds to being infrequent. Paradigm relatedness can be 'radical or transformational': much harder to judge (see [Brown 2014]).

There's a need to be careful as "originality" has been applied to people, not ideas: i.e., the ability of an individual to produce uncommon or unique responses. For example, in Jansson & Smith's [1991] originality metric, it calculates the average "o score" for that person's ideas (average rarity) and divides it by the number of ideas that person had.

7. Conclusions

There are many possible conclusions, just as there are many possible variations in calculations. The comparison set can be limited to set of designs generated by participants in experiment, or have a larger scope limited by expert knowledge. Comparisons done at different "levels": by attributes or functions (i.e., only part of the concept), or with the undivided concept—although when judged by experts it is easy to argue that even they use a subset of all the attributes for judgment. Comparisons can be done by design stage, varying the abstractness of the target design. The main problems are due to relying on the human ability to decide what is important and on the ability to apply heuristics (e.g., similarity matching). The bottom line is that, despite their usefulness, neither SVS nor those approaches based on SVS appear to be very promising as computational methods.

8. References

S.P. Besemer (2006) *Creating products in the age of design. How to improve your new product ideas!* New Forums Press, Inc.

D.C. Brown (2012) "Creativity, Surprise and Design: an introduction and investigation", *Proc. 2nd Int. Conf. on Design Creativity (ICDC2012).*

D.C. Brown (2013) "Developing Computational Design Creativity Systems", *Int. Jnl. of Design Creativity and Innovation 1(1),* 43-55.

D.C. Brown (2014) "Let's not get too creative!", *Proc. Computable Design Creativity Metrics Workshop, 6th Int. Conf. on Design Computing and Cognition (DCC'14)*, London, UK.

C. Charyton, R.J. Jagacinski & J.A. Merrill (2008) "CEDA: A research instrument for creative engineering design assessment", *Psychology of Aesthetics, Creativity, and the Arts, 2(3),* 147-154.

D.H. Cropley & J.C. Kaufman (2012) "Measuring Functional Creativity: Non-Expert Raters and the Creative Solution Diagnosis Scale", *Journal of Creative Behavior 46(2),* 119-137.

D.L. Dean, J.M. Hender, T.L. Rodgers & E.L. Santanen (2006) "Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation", *Jnl. Assoc. for Information Systems 7(10).*

D.G. Jansson & S.M. Smith (1991) "Design fixation", *Design Studies 12(1),* 3-11.

M.L. Maher & D.H. Fisher (2012) "Using AI to Evaluate Creative Designs", *Proc. 2nd Int. Conf. on Design Creativity (ICDC2012),* pp. 45-54.

B.A. Nelson, J. Yen, J.O. Wilson & D. Rosen (2009) "Refined Metrics for Measuring Ideation Effectiveness", *Design Studies 30,* 737-743.

G. Pahl & W. Beitz (1996) *Engineering Design: A Systematic Approach*, Springer.

J. Peeters, P.-A. Verhaegen, D. Vandevenne & J.R. Duflou (2010) "Refined Metrics for Measuring Novelty in Ideation", *Proc. IDMME Virtual Concept 2010.*

P. Sarkar & A. Chakrabarti (2011) "Assessing design creativity", *Design Studies, 32,* 348-383.

M.N. Saunders, C.C. Seepersad & K. Hölttä-Otto (2009) "The Characteristics of Innovative, Mechanical Products", *Proc. ASME 2009 IDETC/CIE Confs.,* DETC2009-87382.

J.J. Shah, N. Vargas-Hernandez & S.M. Smith (2003) "Metrics for measuring ideation effectiveness", *Design Studies, 24*, 111-134.

R. Srivathsavai, N. Genco, K. Hölttä-Otto & C.C. Seepersad (2010) "Study of Existing Metrics Used in Measurement of Ideation Effectiveness", *Proc. ASME 2010 IDETC/CIE Confs.,* DETC2010-28802, Montreal, Canada.

P.-A. Verhaegen, D. Vandevenne & J.R. Duflou (2012) "Originality and Novelty: a different universe", *Proc.DESIGN 2012, the 12th International Design Conference*, Dubrovnik, Croatia, pp. 1961-1966.