Finding Unexpected Patterns in Citizen Science Contributions Using Innovation Analytics

MARY LOU MAHER, MOHAMMAD JAVAD MAHZOON, University of North Carolina at Charlotte

1. INTRODUCTION

Citizen science projects scale up the collection or interpretation of scientific data through a kind of collective intelligence, encouraging non-scientists to participate in scientific challenges. With this often comes a significant increase in data quantity and diversity and along with it the increasing difficulty in sense making that is typical of big data projects. Using exploratory data analysis (Tukey 1977), we can extract and search for different analytical characteristics of data, in which each expresses data in form of a hypothesis. Having massive amounts of data, or in the case of citizen science, very large numbers of contributions, searching for new hypotheses requires significant expertise and is largely guided by heuristics and expectations. Here we address this problem by using innovation analytics to direct the search by looking for *interesting patterns* in the stream of data, which may lead to new and creative hypotheses. Innovation analytics is a framework for an automated search to identify patterns in data that are *unexpected* using computational models of creativity (Grace et al. 2014c).

Research in citizen science data collection applies a variety of data analysis techniques to answer specific questions. For example, Counter et al. (2013) discovered weekend bias in reporting data by citizen scientists. They used statistical methods to correlate citizen scientists' reported data with other phenology resources to find bias in reporting data: Observers tend to report first migratory arrivals on weekends than on weekdays. As another example in the same community, Yu et al. (2014) proposed a model to find misidentified bird species by citizen scientists. They use statistical models to learn species distributions based on environmental features, and detected when users were likely to have misidentified one species as another. Researchers and domain-experts in these communities use their prior knowledge in the domain to decide what structure within the data is interesting and then build models to confirm their hypotheses. However, with innovation analytics we propose to let the data speak for itself: models of unexpectedness will direct the search for structure within the data, creating hypotheses that can be investigated further without requiring prior domain knowledge.

2. INNOVATION ANALYTICS AND DESIGN CREATIVITY

Our current work in innovation analytics builds models of data inspired by computational creativity. Our premise is that creative designs are unexpected and we can use computational models of novelty and expectation to identify unexpected designs. In our previous work, we generated multiple predictive models of design data to identify different kinds of expectations (Grace and Maher 2014). Grace et. al (2014a and 2014b) used regression models and conceptual clustering to build expectations over features of the designs presented in temporal order. By building predictive models from the design data, we can capture trends of data and identify when a new design does not follow the trend. We propose that this approach to data analytics can be applied to citizen science data to identify new and unexpected hypotheses about future scientific data.

To demonstrate the transfer of our innovation analytics approach from identifying creative designs to identifying new hypotheses, we build models of design data and citizen science data and compare the data signatures. For the design data, we use a dataset with about 4000 mobile devices described by several features including width, height, screen size, depth, CPU speed, and RAM size. We build patterns from the data using the Kohonen Self Organizing Map (SOM) (Kohonen 1990) which allows us to reduce multi-dimensional data to two-dimensional maps and produces a pattern signature for

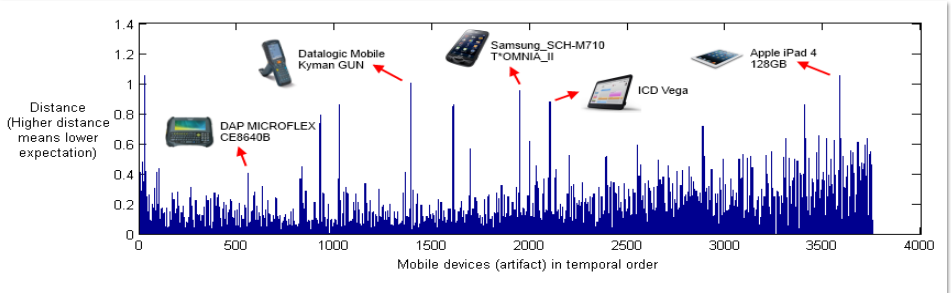

Fig. 1. Pattern signature for mobile devices dataset generated by the SOM. The X axis shows the set of individual artifacts presented to the model in temporal order, and the Y axis is the distance between the artifacts presented to model and the best matching unit in SOM. The distance measure can be interpreted as expectation of the model about the artifact, where a small distance indicates expectation and a large distance is an indication of unexpectedness.

the data. Figure 1 shows the patterns for the mobile devices dataset. Unexpected designs are those at the peaks in the pattern signature. Figure 2 illustrates how representations generated by the SOM can reveal the underlying clusters of designs, which are separated by high distance connections between units in SOM (shown as regions of darker colors).

We use SOM as a novelty detector as has been previously done by Saunders and Gero (2001) and Hodge and Austin (2004). We calculate the distance between a given input and the best matching unit of the SOM (see equation 15 in Hodge and Austin 2004):

$$d = \min\left(\sum_{i=0}^{n-1}\left(x_i(t) - w_{im}(t)\right)\right)$$

Where x is the input data with n features, w is weights of the SOM with m neurons, and t is the time which input x was presented to the model. A close distance means the input is similar to previously learned artifacts, therefore it is expected and not novel, but higher distances indicate an unexpected input and causes the SOM to update its weights. We can measure the change in weights of the SOM in a time period τ with the following formula:

$$\Delta w_\tau = \left(\sum_{i=1}^{m}\left(w_{.i}(t+\tau) - w_{.i}(t)\right)^2\right)/m$$

From the distance graph and Δw_τ we can generalize two kinds of unexpected data items:

(1) Outlier – An outlier occurs in a pattern in which an unexpected data item is not repeated. In the case of an outlier, Δw_τ is not large, but d is relatively large. We use Grubbs' outlier test (1950) to test the significance of the difference between Δw_τ and d. In a design context, outliers are likely to be designs that did not set trends for other devices or were not considered valuable by the consumer. The Datalogic mobile device identified in Figure 1 is an outlier in our mobile devices dataset. This design was primarily used as retail barcode scanners and was not influential in the popular mobile device community.

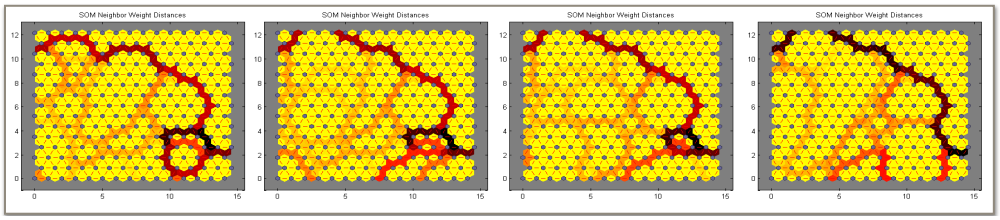

Fig. 2. Representation of the clusters of designs for 40 iterations (the plot was updated after presenting 10 designs to the model). Each plot shows all the units of SOM model in dots (15*15 units arranged in hexagonal grid). Connections between units in the grid are shown with lines, and the background color of each line shows intensity of weight (distance) of that connection. The darker the color is the higher is the weight (distance).

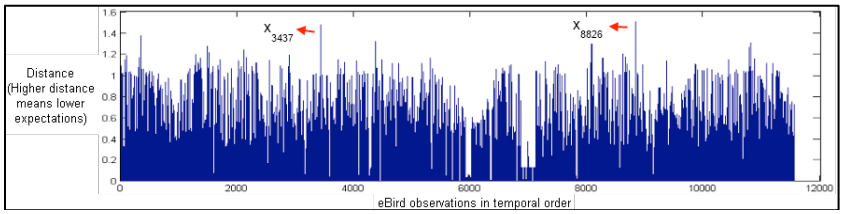

Figure 3: Pattern signature for eBird dataset generated by the proposed method.

	Distance	Taxonomy	Location	Latitude	Longitude	Duration (min)	Effort Dist(km)	Unexpectedness
x_{3434}	0.406419209	14589	US-AK	56.81215	-132.95772	60	3.219	
x_{3435}	0.68218709	30553	US-CA	33.9698513	-117.3194612	75	2.414	
x_{3436}	0.513036798	30553	US-CA	33.957386	-117.3492622	60	1.448	
x_{3437}	1.474137117	21415	US-CA	33.3400894	-116.6445923	1440	1609.34	Harbinger
x_{3438}	0.543984233	4026	US-CA	33.3400894	-116.6445923	1440	1609.34	
x_{3439}	0.28467747	30515	US-CA	33.3400894	-116.6445923	1440	1609.34	
x_{3440}	0.10645686	7429	US-CA	33.3400894	-116.6445923	1440	1609.34	
x_{8823}	0.47590863	215	US-OK	34.7343564	-95.1450455	60		
x_{8824}	0.421072045	14529	US-TX	30.5985609	-103.8923299	109	2.961	
x_{8825}	0.138151278	4492	US-VT	44.5934815	-73.3117418	45	1.609	
x_{8826}	1.500137727	7583.5	US-WA	48.8861942	-122.7861857	1440		Outlier
x_{8827}	0.178876093	14576	US-CA	33.89873	-118.41236	35	0.322	
x_{8828}	0.210130088	3006	US-CA	33.720913	-115.3629684			
x_{8829}	0.238971089	8097	US-CA	34.2470358	-119.1667378	127	0.805	

Table 1: Detected outlier and harbinger for eBird dataset.

(2) Harbinger – A harbinger is an outlier in a pattern in which an unexpected data item is followed by a set of similar data items. Harbingers often signal creation of new trends or styles. In case of a harbinger, d is relatively large, but unlike the outliers Δw_τ is also large. In the design context, these unexpected patterns happen when a new trend is introduced. An example of this in the mobile devices dataset is the occurrence of big display mobiles, which was unexpected when first introduced (ICD Vega in Figure 1), but after a while next generation phones copied this feature.

3. FINDING PATTERNS IN CITIZEN SCIENCE DATA

We applied our predicitve model to citizen science data from the eBird project (ebird.org), a citizen science community that collects bird observations worldwide. As shown in Figure 3 we produced a pattern signature using the SOM model, and identified one outlier and one harbinger in the dataset. Table 1 provides more information about the outlier and harbinger in the highlighted rows. The outlier is a bird observation that has an unexpectedly high duration considering the location and the bird taxonomy of the observation. The harbinger is a bird observation that has an unexpectedly high duration and effort distance given its location and taxonomy, but is followed by a set of similar observations. This harbinger may be an indication of a new trend, and signal the possibility of a new hypothesis.

4. SUMMARY

We have presented a comparison of pattern signatures of design data and citizen science data using a self orgaizing map algorithm to generate the patterns. Inspired by computatoinal models of creativity, we recognize that an unexpected design has the potential to be creative. In our pattern signatures of design data, we recognized the difference between an outlier and a harbinger. Where an outlier is a creative design that did not cause a new trend, a harbinger was the first of a new kind of design. We use this simple interpretation to explore the possibilty of an analytic approach to identifying potential hypotheses in citizen science data. An outlier in citizen science data may occur when the person recording the observation made errors in data collection or reporting. A harbinger in citizen science data may be an unexpected trend and direct the scientist to generating a new hypothesis from the data. We are continuing to pursue the use of unsupervised learning to identify patterns in citizen science data to facilitate the ability to recognize when an unexpected data item is low quality data or a new hypothesis.

REFERENCES

Courter, J. R., Johnson, R. J., Stuyck, C. M., Lang, B. A., & Kaiser, E. W. (2013). Weekend bias in Citizen Science data reporting: implications for phenology studies. International journal of biometeorology, 57(5), 715-720.

eBird Basic Dataset. Version: EBD_relFeb-2014. www.ebird.org, Cornell Lab of Ornithology, Ithaca, New York. (last accessed July 2014).

Grace, K., Maher, M. L., Fisher, D., & Brady, K. (2014a). Data-intensive evaluation of design creativity using novelty, value, and surprise. International Journal of Design Creativity and Innovation, (ahead-of-print), 1-23.

Grace, K., Maher, M. L., Fisher, D., & Brady, K. (2014b). Modeling Expectation for Evaluating Surprise in Design Creativity. Paper presented at the DCC'14: International Conference on Design Computing and Cognition, London, UK.

Grace, K., & Maher, M. L. (2014). Using Computational Creativity to Guide Data-Intensive Scientific Discovery. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Grubbs, Frank E. (1950). Sample criteria for testing outlying observations. The Annals of Mathematical Statistics, 27-58.

Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. Artificial Intelligence Review, 22(2), 85-126.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464-1480.

Saunders, R., & Gero, J. S. (2001). A curious design agent. In *CAADRIA* (Vol. 1, pp. 345-350).

Tukey, J. W. (1977). Exploratory data analysis. *Reading, Ma*, *231*, 32.

Yu, J., Hutchinson, R. A., & Wong, W. K. (2014, June). A Latent Variable Model for Discovering Bird Species Commonly Misidentified by Citizen Scientists. In Twenty-Eighth AAAI Conference on Artificial Intelligence.

ACKNOWLEDGMENTS

This project is funded by the National Science Foundation in a grant to UNC Charlotte, the University of Maryland, and the University of Colorado. The award is titled: "EAGER: Collaborative Research: A Computational Model for Evaluating the Quality of Citizen Science Contributions". The authors acknowledge that many of the ideas in this paper are based on discussions with Kazjon Grace, Tom Yeh, Jennifer Preece, and Carol Boston.