Proceedings of Informing Science & IT Education Conference (InSITE) 2015

Cite as: Scime, A. (2015). A simple measure of diversity. *Proceedings of Informing Science & IT Education Conference (InSITE) 2015,* 367-383. Retrieved from http://Proceedings.InformingScience.org/InSITE2015/InSITE15p367-383Scime1503.pdf

A Simple Measure of Diversity

Anthony Scime
The College at Brockport, State University of New York, Brockport, NY, USA

ascime@brockport.edu

Abstract

Diversity is a relative concept, which has been applied to a number of domains, such as culture and biology. A simple measure of diversity is derived by drawing from the domains of biological and cultural diversity, as well as, information retrieval for its measurement capabilities. This domain independent diversity measure can be used to determine diversity between entities in any collection that can be expressed as features and their values. The measurement can be applied to a single feature or to any combination of features. The entities may be, among other things, words in a document, biological species in an environment, people in an organization, or records in a data set. This diversity measure provides a single value for entities in their collection; measuring the relative diversity of the entity with respect to the other entities in the collection. This tool can be used to compare and contrast diversity between collections of entities, or within the same collection over time.

Keywords: Biological Diversity, Cultural Diversity, Diversity, Information Retrieval, Measurement

Introduction

Diversity is what makes life interesting. It is an important concept in understanding both the physical and cultural worlds. Diversity exists in every collection of entities in a domain. The domains may be biological, cultural, geological, textual, or informational. The entities may be organisms, from microbes to people, organizations, rocks, the words in a document, or data in a data set.

Diversity understanding leads to important decisions. Cultural diversity understanding is important for decision and policy making, to ensure equality, prevent discrimination and promote inclusiveness, which in turn lead to peace and harmony (Jones, 1994; Trickett, Watts, & Birman, 1994). In biology an understanding of a biological collection's diversity helps in establishing policies and actions that protect and conserve the environment (Duffy, 2009).

While we know diversity is important, there needs to be a method to determine a collection's level of diversity. Once a measurement is achieved collections can be compared on diversity. This comparison may be of the collection at different periods of time or between different, but similar, collections. Measurement allows evaluation of the effect of an action on a collection's diversity.

Editor: Eli Cohen

In biology diversity is considered little more than the effective number of species present in the current study (Hill, 1973; Magurran, 2004). Culturally diversity is considered the differences related to social class, ethnicity, culture, and language between a group of two or more people (McGrath, Berdahl, & Arrow, 1995; Trickett, Watts, & Birman, 1994; Zeichner, 1993). More generally, diversity is the quality or state of having different forms, types, ideas, etc. (Merriam-Webster, 2014). Diversity then is the difference within a collection of entities based on some common features of the entities. These basis features have a range of values that establish the characteristics of each entity. Diversity is the difference between these characteristics.

Cultural diversity has been a state to which organizations have been striving for a number of years. There is a common belief that there is value in organizational diversity (Cox, Lobel, & McLeod, 1991; Hoffman, 1959; Hoffman & Maier, 1961). When an organization contains people that have a range of knowledge, experience, and perspectives they as a group will be able to make better, more effective decisions. For example, teams diverse with respect to educational specialization are more willing to accept and effect change (Wiersema & Bantel, 1992).

Biologically, the effect of diversity is predictable. It is necessary to sustain life. Different species in an environment affects the ecosystem, which then impacts human life. Combinations of species in an ecosystem create more biomass and use more resources than do a single species. This is true across taxa, trophic levels, and habitats (Duffy, 2009). The availability of diverse food sources creates growth of the predator population, which in turn has a positive effect on the food chain (Duffy et al., 2007). Plant diversity positively controls productivity and nutrient use while protecting against disease and climate change (Loreau et al., 2001). At the microcosm level similar benefits are found (Naeem & Li, 1997).

How diversity is measured varies with the domain. Cultural diversity measurement is problematic. There are two basic approaches to measuring organizational cultural diversity (Mannix & Neale, 2005). Factor approaches are where differences are measured and compared for two or more different features. These may include visible features like race, gender, age; or underlying features like employee turnover, minorities in executive positions, cost per hire, penetration of minority markets, recycle time, resolution of customer issues, or any other items related to aspects of an organization's operation. Typically, these factors are evaluated and compared subjectively on a diversity scorecard (Hubbard, 2004). A scorecard is simply a list of factors and their values. But, scorecards do not provide a final single value for easy comparison (Jensen, 2001).

The second approach to organizational cultural diversity measurement is a proportional approach. Kanter (1977) defines in-group and out-group entities based on an identifiable feature. The group is defined as one of four types. A uniform group is when the value of the feature is the same for all entities. A skewed group has a minority membership of between 1% and 15%. A tilted group has a minority of 15% to 35%. A range of 35% to 65% constitutes a balanced group. These percentages of group definitions are somewhat arbitrary, and are based on measurement of a single factor.

At a larger cultural level, Hofstede (2001a) developed the standard factors (or dimensions) of power distance, uncertainty avoidance, individualism, masculinity versus femininity, and long-tern versus short-term orientation that have been used to measure and compare national and organizational cultures. These dimensions were developed using statistical trends across 40 countries. These dimensions have been used extensity and extended by social science researchers to explain differences between countries and companies. The results can be expressed as index scores for comparison. A recognized problem with some studies using this model is that the methods are limited to the pre-established dimensions and the questions used as an instrument are sometimes mis-written or mis-applied (Hofstede, 2001b; Minkov & Hofstede, 2014).

Biodiversity has a number of different methods of measurement, which depend on the biodiversity being investigated; among these are functional diversity, taxonomic diversity, and phylogenetic (evolutionary) diversity (Petchey & Gaston, 2002). Leinster and Cobbold (2012) have summarized and unified the most commonly used diversity measures of biology. These biological diversity measures depend on two factors: relative abundance and similarity. Relative abundance is the proportion of each species present in the community under investigation. Similarity is a measure of biological distance between species in the community. Additionally, these measures may contain a sensitivity parameter that controls the relative importance the measurement user places on rare species.

Both cultural and biological diversity experts recognize that diversity is relative ((Hofstede, 2001a; Kanter, 1977; Leinster & Cobbold, 2012). The characteristics on which diversity is based is dependent on the purpose of the measurement. Within a collection of entities, some entities may be the same, some are different from one another, and, commonly, some differ in only some features. An entity in a collection where all the entities are the same is not diverse. A collection where every entity is totally different from every other entity is totally diverse, and all the entities are equally diverse.

Biodiversity and cultural diversity measures consider the diversity of a collection. Because diversity is relative, changing a collection's diversity is a function of the addition or subtraction of the appropriate individual entities. Yet, biodiversity and cultural diversity measures do not measure the diversity of the individual entities. The entity within a collection that is the most different from the other entities in the collection is the most diverse with respect to the evaluated characteristics and has the most effect on the collection's diversity.

Drawing from biological and cultural diversify as well as information retrieval for its measurement capabilities (as seen below) a simple measure of diversity is derived based on entity features and their values in a collection. This simple measure of diversity is applicable to biological and cultural systems or any collection which can be defined in terms of definable features and values. That is, any collection on which data can be obtained.

This research proceeds as follows; first a sample generic collection of entities is presented that is subsequently used to demonstrate measuring diversity. This is followed by the development of the diversity measure based on work in information retrieval. The paper then proceeds to use this measurement method with examples taken from the sample collection to show how the method measures diversity. This includes demonstrating how the measurement can be applied between collections, as well as within a collection. Limitations and future work are discussed next, followed by the conclusion section, which discusses the relationship of this method to biological and cultural diversity, as well as future applications.

A Generic Collection

A collection of entities exists with three levels – the collection level, the entity level, and the characteristic level. A characteristic is a feature and its value for that entity. When considering diversity the entities are diverse from each other within the collection considering certain characteristics. An entity may have many characteristics and the level of diversity of an entity is relative to the collection in which it resides.

Consider a set of figures. Each figure has four features – shape (triangle, rectangle, star, arrow), shade (black, grey, white), size (small, medium, large), and orientation (upright, tilted). There are then 72 different combinations of features (Figure 1). Each figure is unique having a different combination of shape, shade, size, and orientation. But, it is possible for any figure to be replicated, to occur more than once in a collection.

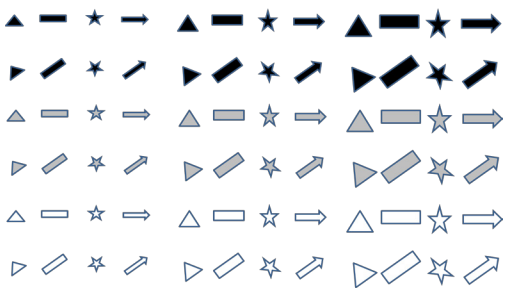

Figure 1: The set of all figures

A figure in a collection is diverse relative to the other figures if it is different from the other figures with respect to the basis for determining diversity. The basis is the features. The diversity of a figure is a measure of difference. It is a relative measure with respect to the other figures in the same collection. A diverse collection is a collection in which the figures are different from each other.

Examples of collections are the homogeneous and the totally diverse. Select from the universe of figures 10 medium gray upright arrows (Figure 2.). This collection is homogeneous and has no diversity. Also, possible is the opposite situation, a collection where every figure is different (Figure 3), that is, totally diverse.

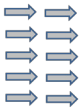

Figure 2: Collection of 10 medium gray upright arrows

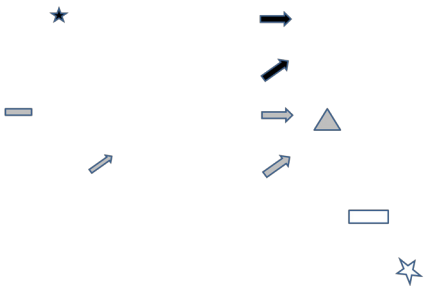

Figure 3: Collection of 10 different figures

For a collection where some of the figures have some characteristics in common a measurement method is needed to determine diversity. The next section develops a simple measurement method for diversity from information retrieval concepts.

Measuring Diversity

In terms of a random but meaningful collection of entities consider a written document. The entities in a document are the words, which are placed in an order to convey a meaning to the reader. Yet, as a collection of words or terms, where the order does not matter and any word can occur multiple times, a document is a random collection of entities.

Salton and Yang (1973) investigated document retrieval as a function of term frequency and term frequency distribution in a document collection. The objective in document retrieval is to find all the documents that contain a term (recall) and only those documents that contain the term (precision). The occurrence of a term in all documents does not facilitate distinguishing the document from the others in the collection, and thus does not help in retrieval; those are words that are too common. Terms with a medium level of frequency and a reasonably shewed distribution should provide acceptable levels of both recall and precision. Precision is perhaps best when term frequency is very skewed. That is, the term occurs in very few documents.

If our collection is considered similar to a document the figures are the terms. The measure of a figure's diversity is a function of the uncommonness of the figure in the collection with respect to its characteristics. This is in keeping with Sparck Jones' (2004) concept of inverse document frequency, where terms with low frequency in a document are stressed, and precision in document retrieval is improved.

Diversity is a measure of the distinctness of the combination of characteristics in an entity with respect to the other entities in the collection. Borrowing from information retrieval, we define a measure of diversity for an entity, the entity diversity measure (D_E), as the sum of the inverses of commonness of an entity's characteristics. Because each characteristic (i) occurs only once in an entity; the inverse of commonness of occurrences (C_i) is the inverse of the count of the characteristic's occurrences in the collection (O_i).

$$C_i = \begin{cases} 1/O_i \; if \; the \; charateristic \; exists \; in \; the \; figure \\ 0 \; otherwise \end{cases} \quad (1)$$

$$D_E = \sum C_i \quad (2)$$

Where:

O_i = Count of the occurrences of characteristic i in the collection

C_i = Inverse of commonness of characteristic i in the figure

The diversity measure may be normalized (D_{Norm}) between 0 and 1 to provide easy comparison within a collection.

$$D_{Norm} = \frac{D_E}{MaxD_E} \quad (3)$$

Care should be taken when all the entities in a collection have a $D_{Norm} = 1$. This occurs when the entities are all different (maximum diversity) and when all the entities are the same (no diversity). A normalized value of one and no diversity occurs because diversity is relative. When there is no difference between the entities there is no basis for comparing characteristics. When all the normalized values are one the diversity measure values need to be considered to determine if there is diversity. This measure of diversity is now demonstrated using subsets of the generic collection.

Entity Diversity within a Collection

Applying the measure to Figure 2, where all the entities, the figures, in the collection are the same, finds that all the figures have the same diversity measure. Table 1 provided the count of occurrences for each arrow and its inverse. Table 2 calculates the diversity measure for an arrow, which is the same for each medium gray upright arrow. Each figure has the diversity measure value of 0.4, normalized to 1. Because the normalized values are one and the diversity measure values are less than one, this collection has no diversity.

Table 1: Occurrences of characteristics for each medium gray upright arrow

CHARACTERISTIC	O_i	$1/ O_i$
Size = Medium	10	0.1
Color = Gray	10	0.1
Orientation = Upright	10	0.1
Shape = Arrow	10	0.1

Table 2: Sum of $1/O_i$ for each arrow

FIGURE	ΣC_i	D_E	D_{Norm}
Medium Gray Upright Arrow	0.1+0.1+0.1+0.1	0.400	1.000

Now consider the collection of figures in Figure 3, where all the figures are different. We calculate a diversity measure value for each figure. Table 3 provides the number of occurrences of each characteristic and its inverse. Table 4 calculates the diversity measure for each figure. The large gray upright triangle is the most diverse figure and the medium gray upright arrow is the least diverse.

Table 3: Occurrences of characteristics for figures in figure 3		
CHARACTERISTIC	**O_i**	**$1/O_i$**
Size = Small	3	0.333
Size = Medium	4	0.250
Size = Large	3	0.333
Color = Black	3	0.333
Color = Gray	5	0.200
Color = White	2	0.500
Orientation = Upright	7	0.143
Orientation = Tilted	3	0.333
Shape = Triangle	1	1.000
Shape = Rectangle	2	0.500
Shape = Arrow	5	0.200
Shape = Star	2	0.500

Table 4: Diversity measure values for figures in figure 3			
FIGURE	**ΣC_i**	**D_E**	**D_{Norm}**
Small Black Upright Star	0.333+0+0+0.333+0+0+0.143+0+0+0+0+0.500	1.309	0.781
Medium Black Upright Arrow	0+0.250+0+0.333+0+0+0.143+0+0+0+0.200+0	0.926	0.553
Medium Black Tilted Arrow	0+0.250+0+0.333+0+0+0+0.333+0+0+0.200+0	1.116	0.666
Small Gray Upright Rectangle	0.333+0+0+0+0.200+0+0.143+0+0+0.500+0+0	1.176	0.702
Medium Gray Upright Arrow	0+0.250+0+0+0.200+0+0.143+0+0+0+0.200+0	0.793	0.473
Large Gray Upright Triangle	0+0+0.333+0+0.200+0+0.143+0+1.000+0+0+0	1.676	1.000
Small Gray Tilted Arrow	0.333+0+0+0+0.200+0+0+0.333+0+0+0.200+0	1.066	0.636
Medium Gray Tilted Arrow	0+0.250+0+0+0.200+0+0+0.333+0+0+0.200+0	0.983	0.587
Large White Upright Rectangle	0+0+0.333+0+0+0.500+0.143+0+0+0.500+0+0	1.476	0.881
Large White Tilted Star	0+0+0.333+0+0+0.500+0+0.333+0+0+0+0.500	1.666	0.994

It is possible that some figures could have all the same characteristics as another figure. Duplicate the large gray upright triangle and remove the medium black tilted arrow to create another collection (Figure 4).

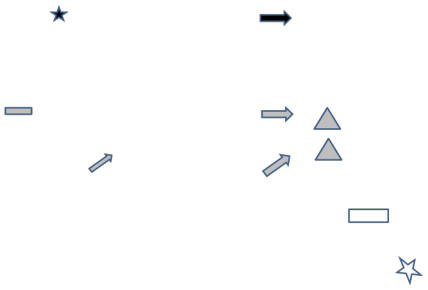

Figure 4: Collection with one figure the same

Recalculating the figures' diversity measures is in done in Tables 5 and 6. The diversity value changed for every figure because the changes affect every figure. The large white tilted star is now the most diverse. The least diverse is still the medium gray upright arrow.

Eliminating the medium black tilted arrow and adding a large gray upright triangle changes the diversity measure value of every figure. Even with two large gray upright triangles, because of the combination and commonness of characteristics the large gray upright triangles, which was the most diverse previously, do not become the least diverse.

Table 5: Occurrences of characteristics for figures in figure 4

CHARACTERISTIC	O_i	$1/O_i$
Size = Small	3	0.333
Size = Medium	3	0.333
Size = Large	4	0.250
Color = Black	2	0.500
Color = Gray	6	0.167
Color = White	2	0.500
Orientation = Upright	7	0.143
Orientation = Tilted	3	0.333
Shape = Triangle	2	0.500
Shape = Rectangle	2	0.500
Shape = Arrow	4	0.250
Shape = Star	2	0.500

Table 6. Diversity measure values for figures in figure 4			
FIGURE	**ΣC_i**	**D_E**	**D_{Norm}**
Small Black Upright Star	0.333+0+0+0.500+0+0+0.143+0+0+0+0+0.500	1.476	0.932
Medium Black Upright Arrow	0+0.333+0+0.500+0+0+0.143+0+0+0+0.250+0	1.226	0.774
Medium Black Tilted Arrow	0+0.333+0+0.500+0+0+0+0.333+0+0+0.250+0	1.416	0.895
Small Gray Upright Rectangle	0.333+0+0+0+0.167+0+0.143+0+0+0.500+0+0	1.143	0.722
Medium Gray Upright Arrow	0+0.333+0+0+0.167+0+0.143+0+0+0+0.250+0	0.893	0.564
Large Gray Upright Triangle 1	0+0+0.250+0+0.167+0+0.143+0+0.500+0+0+0	1.060	0.670
Small Gray Tilted Arrow	0.333+0+0+0+0.167+0+0+0.333+0+0+0.250+0	1.083	0.684
Large Gray Upright Triangle 2	0+0+0.250+0+0.167+0+0.143+0+0.500+0+0+0	1.060	0.670
Large White Upright Rectangle	0+0+0.250+0+0+0.500+0.143+0+0+0.500+0+0	1.393	0.880
Large White Tilted Star	0+0+0.250+0+0+0.500+0+0.333+0+0+0+0.500	1.583	1.000

It may be desirable to measure diversity based on only one feature. Such a measure removes the interplay effects of the characteristics, such as seen above where the large gray upright triangle did not become the least diverse. Consider a collection of four figures differing in only the shape feature (Figure 5). Tables 7 and 8 calculate the diversity measure values. Each figure is different than the others, but they all have the same diversity measure value in this collection. The figures are equally diverse from each other. The diversity measure values are greater than or equal to one, therefore the collection is diverse. If the calculation is done only on the shape feature, the diversity measure value changes, and the normalized value (D_{Norm}) remains one (Tables 9 and 10).

However, if shape is not a feature to be considered the figures are still found to have the same diversity measure values (Tables 11 and 12). As in the first example, the normalized values are one and the diversity measure values are less than one, this collection has no diversity when shape is not a consideration.

Figure 5: Collection of four figures differing in one feature

Table 7: Occurrences of characteristics for figures in figure 5		
CHARACTERISTIC	**O_i**	**$1/O_i$**
Size = Small	4	0.25
Color = Black	4	0.25
Orientation = Upright	4	0.25
Shape = Triangle	1	1.0
Shape = Rectangle	1	1.0
Shape = Star	1	1.0
Shape = Arrow	1	1.0

Table 8: Diversity measure values for figures in figure 5			
FIGURE	**ΣC_i**	**D_E**	**D_{Norm}**
Small Black Upright Triangle	0.25+0.25+0.25+1+0+0+0	1.75	1.000
Small Black Upright Rectangle	0.25+0.25+0.25+0+1+0+0	1.75	1.000
Small Black Upright Star	0.25+0.25+0.25+0+0+1+0	1.75	1.000
Small Black Upright Arrow	0.25+0.25+0.25+0+0+0+1	1.75	1.000

Table 9: Occurrences of characteristics for shape for figures in figure 5		
CHARACTERISTIC	**O_i**	**$1/O_i$**
Shape = Triangle	1	1.0
Shape = Rectangle	1	1.0
Shape = Star	1	1.0
Shape = Arrow	1	1.0

Table 10: Diversity measure values for figures in figure 5 considering only shape			
FIGURE - SHAPE	**ΣC_i**	**D_E**	**D_{Norm}**
Triangle	1	1	1.000
Rectangle	1	1	1.000
Star	1	1	1.000
Arrow	1	1	1.000

Table 11: Occurrences of characteristics other than shape for figures in figure 5		
CHARACTERISTIC	**O_i**	**$1/O_i$**
Size = Small	4	0.25
Color = Black	4	0.25
Orientation = Upright	4	0.25

Table 12: Diversity measure values for figures not considering shape in figure 5			
FIGURE	**ΣC_i**	**D_E**	**D_{Norm}**
Small Black Upright	0.25+0.25+0.25	0.75	1.000

Thus far we have established the concept of an entity having a diversity measure value within a collection, and that this value can be normalized to simplify comparison within the collection. Because the diversity measure is relative, the collection itself does not have a diversity measure value until it is compared to another collection. Diversity can also be determined between collections as shown in the next section.

Collection Diversity between Collections

To determine the relative diverseness between multiple collections the same calculations are performed. A group of collections still has three levels – the group, the collections, and the entities, in our case the figures. Calculating the diversity measure for each collection (Figures 2, 3 and 4) in the group based on the figures is shown in Tables 13 and 14. As would be expected the collection with all the figures different is the most diverse and the collection of all Medium Gray Upright Arrows is the least diverse. The collection with one different figure is much closer in value to the ten different figure collection than the same figure collection, as would be expected.

Table 13: Occurrences of figures in the group of 3 collections		
FIGURE	**O_i**	**$1/O_i$**
Small Black Upright Star	2	0.500
Medium Black Upright Arrow	2	0.500
Medium Black Tilted Arrow	2	0.500
Small Gray Upright Rectangle	2	0.500
Medium Gray Upright Arrow	13	0.077
Large Gray Upright Triangle	2	0.500
Small Gray Tilted Arrow	2	0.500
Medium Gray Tilted Arrow	1	1.000
Large White Upright Rectangle	2	0.500
Large White Tilted Star	2	0.500

Table 14: Diversity measure values for collections in the group of 3 collections			
COLLECTION	**ΣC_i**	**D_E**	**D_{Norm}**
The Same Figures (Figure 2)	0+0+0+0+0.077+0+0+0+0+0	0.077	0.015
Ten Different Figures (Figure 3)	0.5+0.5+0.5+0.5+0.077+0.5+0.5+1.0+0.5+0.5	5.077	1.000
One Figure the Same (Figure 4)	0.5+0.5+0.5+0.5+0.077+0.5+0.5+0+0.5+0.5	4.077	0.803

Consider the case where the collections sizes are not the same. Clearly, a collection with different figures should be more diverse than a homogeneous collection, regardless of size. Figure 6 shows a diverse collection of five figures all different and Tables 15 and 16 calculate the diversity measure values for a collection's group consisting of the collections in Figures 2 and 6. As expected, the collection of five different figures is significantly more diverse.

Figure 6: Collection with 5 different figures

Table 15: Occurrences of figures in the group of 2 collections		
FIGURE	**O_i**	**$1/O_i$**
Small Black Upright Star	1	1.000
Medium Black Tilted Arrow	1	1.000
Medium Gray Upright Arrow	11	0.091
Large Gray Upright Triangle	1	1.000
Large White Tilted Star	1	1.000

Table 16: Diversity measure values for collections in the group of 2 collections			
COLLECTION	**ΣC_i**	**D_E**	**D_{Norm}**
The Same Ten Figures (Figure 2)	0+0+0.091+0+0	0.091	0.022
Five Different Figures (Figure 6)	1.000+1.000+0.091+1.000+1.000	4.091	1.000

Finally consider the reverse, a collection of five figures all the same (Figure 7) compared to Figure 3 with ten different figures. Again as expected, the more diverse collection is the collection with the different figures (Tables 17 and 18).

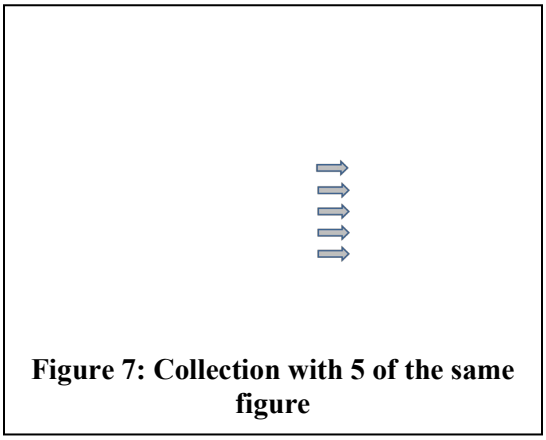

Figure 7: Collection with 5 of the same figure

Table 17: Occurrences of figures in the second group of 2 collections		
FIGURE	**O_i**	**1/Oi**
Small Black Upright Star	1	1.000
Medium Black Upright Arrow	1	1.000
Medium Black Tilted Arrow	1	1.000
Small Gray Upright Rectangle	1	1.000
Medium Gray Upright Arrow	6	0.167
Large Gray Upright Triangle	1	1.000
Small Gray Tilted Arrow	1	1.000
Medium Gray Tilted Arrow	1	1.000
Large White Upright Rectangle	1	1.000
Large White Tilted Star	1	1.000

Table 18: Diversity measure values for collections in the second group of 2 collections			
COLLECTION	**ΣC_i**	**D_E**	**D_{Norm}**
Ten Different Figures (Figure 3)	1+1+1+1+1+0.167+1+1+1+1+1+1	9.167	1.000
Five Figures the Same (Figure 7)	0+0+0+0+0+0.167+0+0+0+0+0+0	0.167	0.018

Thus far demonstrated has been how the measure of diversity can be used in any collection of entities, even when the entities are collections themselves. There are limitations and future work that can proceed from here.

Limitations and Future Work

The shortcoming of this measure is in its normalization. The normalized measure cannot always distinguish completely diverse collections from collections with no diversity. In such cases all the entities in the collection have the same value and normalize to one. However, no diversity in most collections would be very unusual and a simple review of the diversity measure values will indicate whether or not the collection contains entities that are not diverse.

A second problem is when the feature's possible values are not distinct. Such a situation could occur in a data set where the feature is an attribute with a domain of continuous values. Even so, that attribute's value for a given entity in the data set is a single value. Therefore this approach is still possible, because it considers the attribute-value pair (the characteristic) as the basis for evaluation. Attribute-value pairs are counted, regardless of the type of attribute.

But, continuous values could lead to many attribute-value pairs. Future research may find methods to handle these continuous attributes. For example, the number of attribute-value pairs can be reduced by grouping. For attributes with interval and ratio scales, within the attribute itself there is a distance between values. Given attribute A (AttrA) with possible continuous values of 1

through 9, the distance between 3 and 5 is 2 which is less than the distance between 3 and 7 (a distance of 4), or 4 and 9 (a distance of 5). It may be desirable to consider entities with AttrA values close together as the same, and counted as if they were the same value.

Values of AttrA can be grouped together, as for example, Gp1 (values 1, 2, 3), Gp2 (values 4, 5, 6) and Gp3 (7, 8, 9). The group designation (Gp1, Gp2, or Gp3) replacing the values in the entities for AttrA. This then discretizes AttrA's values and the simple measure of diversity approach treats values 1, 2, 3 as the same value; 4, 5, 6 as the same value; and 7, 8, 9 as the same value. How many groups and their components is dependent on the meaning of AttrA, and the purpose of the diversity analysis.

Regardless of these limitations the measure for diversity is applicable to collections or datasets of entities.

Conclusion

Formulated is a measure for determining diversity between entities in a collection. The procedure is summarized in Figure 8. The entities may be words in a document, but more importantly they may be biological species in an environment, people in an organization, or the organization itself. The entities may be the records in any data set, a collection, which consists of attributes and their values; in this case, it is the attribute-value pairs that are the characteristics being counted and compared. This diversity measure provides a value for entities in their collection; measuring the relative diversity of the entity with respect to the other entities in the collection.

To determine an entity's diversity measure:

1. Count the number of occurrences of each characteristic in the collection.
2. Calculate the inverse of the occurrence count for each characteristic.
3. Sum the inverses of each occurrence count that is present in each entity.

To compare diversity measures of entities in a collection:

1. Find the maximum diversity measure value.
2. Divide each entity's diversity measure value by the maximum diversity measure value.

Figure 8: Summary of steps to measure diversity

Unlike a diversity scorecard, this measure provides a single value for each entity that can be used to compare entities in the collection. The factors on the entity scorecard may be useful as features in calculation of the diversity measure. This measure is not limited to established dimensions like those used to evaluate cultural differences between companies or countries. Those established dimensions can be used as the features, and the user of the measure can choose to add other features important to their particular need for measuring diversity.

Like biodiversity measures this measure is relative and subjective. It is relative to the other entities in the collection under consideration, and subjective in the selection of features on which diversity is based. Unlike the biodiversity measures it does not have a difficult to set sensitivity parameter, nor does it require distance measures to be determined between the entities.

As examples of diversity measure use, in organizations cultural diversity measurement is done to determine if the organization is progressing in their efforts to become more diverse. As a decision making tool, diversity measure values can be calculated and compared for changes when an event happens or is contemplated in an organization. If an increase in diversity is desired, the sensitivity of the organizational diversity to a proposed change can be tested before the change is implemented. If the diversity measure value does not improve relative to the existing value, that is, diversity does not improve, the event may be canceled.

Additionally, it is generally accepted that modernization has provided the opportunity for cultural diversity to increase, yet in any particular region differences and, thus diversity, has decreased (Newson, Richerson, & Boyd, 2007). The measurement tool presented provides a method to determine the change in diversity in a region or as a function of modernization over time.

Likewise, in ecosystem policy decisions changes to the environment can have an effect on organisms and eventually humanity. Measuring the change in diversity resulting from policy implications can be an important factor in deciding to proceed with a new policy.

As a final example, in information retrieval and Web search, ranking of retrieval results is often done by popularity. Search engine algorithms measure a Web page's popularity as its authority; a measure of the number of incoming links to the page, the number "clicks" a page has received from others, or other indicators of popularity. But, this authority (popularity) ranking does not lead to discovery of new or different information. Using a diversity measure may lead to identification of search results that are unusual, and perhaps more applicable to the searcher's information need.

References

Cox, T., Lobel, S., & McLeod, P. (1991). Effects of ethnic group cultural differences on cooperative and competitive behavior on a group task. *Academy of Management Journal*, *34*, 827-847.

Duffy, J. E. (2009). Why biodiversity is important to the functioning of real-world ecosystems. *Frontiers in Ecology and the Environment*, *7*(8), 437-444.

Duffy, J. E., Cardinale, B. J., France, K. E., McIntyre, P. B., Thebault, E., & Loreau, M. (2007). The functional role of biodiversity in ecosystems: Incorporating trophic complexity. *Ecology Letters*, *10*, 522-538.

Hill, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology*, *54*(2), 427-432.

Hoffman, L. (1959). Homogeneity and member personality and its effect on group problem solving. *Journal of Abnormal and Social Psychology*, *58*, 27-32.

Hoffman, L., & Maier, N. (1961). Quality and acceptance of problem solutions by members of homogeneous and heterogeneous groups. *Journal of Abnormal and Social Psychology*, *62*, 401-407.

Hofstede, G. (2001a). *Culture's consequences* (2nd ed.). Thousand Oaks, CA: Sage.

Hofstede, G. (2001b). Culture's recent consequences: Using dimension scores in theory and research. *International Journal of Cross-Cultural Management, 1*(1), 11-30.

Hubbard, E. E. (2004). *The diversity scorecard: Evaluating the impact of diversity on organizational performance*. Oxford: Elsevier.

Jensen, M. C. (2001). Value maximisation, stakeholder theory, and the corporate objective function. *European Financial Management*, *7*(3), 297-318.

Jones, J. M. (1994). Our similarities are different: Toward a psychology of affirmative diversity. In E. J. Trickett, R. J. Watts, & D. Birman (Eds.), *Human diversity* (pp, 27-45). San Francisco: Jossey-Bass.

Kanter, R. (1977). Some effects of proportions on group life: Skewed sex ratios and responses to token women. *American Journal of Sociology*, *82*, 965-990.

Leinster, T., & Cobbold, C. A. (2012). Measuring diversity: The importance of species similarity. *Ecology*, *93*, 477-489.

Loreau M., Naeem S., Inchausti P., Bengtsson J., Grime, J. P., Hector, A., et al. (2001). Biodiversity and ecosystem functioning: Current knowledge and future challenges. *Science*, *294*, 804-808.

Magurran, A. E. (2004). *Measuring biological diversity*. Oxford: Wiley-Blackwell.

Mannix, E., & Neale, M. A. (2005). What differences make a difference? The promise and reality of diverse teams in organizations. *Psychological Science in the Public Interest*, *6*(2), 31-55.

McGrath, J. E., Berdahl, J. L., & Arrow, H. (1995). Traits, expectations, culture, and clout: The dynamics of diversity in work groups. In S. E. Jackson & M. N. Ruderman (Eds.), *Diversity in work teams* (pp. 17-45). Washington: American Psychological Association.

Merriam-Webster (2014). *Diversity*. Retrieved July 4, 2014 from http://www.merriam-webster.com/dictionary/diversity

Minkov, M., & Hofstede, G. (2014). A replication of Hofstede's uncertainty avoidance dimenstion across nationally representative samples from Europe. *International Journal of Cross-Cultural Management, 14*(2), 161-171.

Naeem S., & Li, S. (1997). Biodiversity enhances ecosystem predictability. *Nature*, *390*, 507-509.

Newson, L., Richerson, P. J., Boyd, R. (2007). Cultural evolution and the shaping of cultural diversity. In S. Kitayama & D. Cohen (Eds.), *Handbook of cultural psychology* (pp. 454-476). New York: Guilford.

Petchey, O. L., & Gaston, K. J. (2002). Functional diversity (FD), species richness and community composition. *Ecology Letters*, *5*, 402-411.

Salton, G., & Yang, C. S. (1973). On the specification of term values in automatic indexing. *The Journal of Documentation*, *29*(4), 351-372.

Sparck Jones, K. (2004). A statistical interpretation of term specificity and its application to retrieval. *Journal of Documentation*, *60*(5), 493-502.

Trickett, E. J., Watts, R. J., & Birman, D. (1994). Toward an overarching framework for diversity. In E. J. Trickett, R. J. Watts, & D. Birman (Eds.), *Human diversity* (pp. 7-26). San Francisco: Jossey-Bass.

Wiersema, M., & Bantel, K. (1992). Top management team demography and corporate strategic change. *Academy of Management Journal*, *35*, 91-121.

Zeichner, K. M. (1993). *Educating teachers for cultural diversity*. Special Report, National Center for Research on Teacher Learning. Retrieved January 29, 2015 from http://files.eric.ed.gov/fulltext/ED359167.pdf

Biography

Anthony Scime is a graduate of George Mason University with an interdisciplinary doctorate in Information Systems and Education. Currently, he is an Associate Professor of Computer Science at The College at Brockport, State University of New York. His work in data mining has been published in *Expert Systems with Applications,* the *International Journal of Business Intelligence and Data Mining, Data Analysis Techniques and Strategies, Social Sciences Quarterly,* and *Public Opinion Quarterly*. His current research interests include the data mining in the social and behavioral sciences, measures of interestingness, and computing education.