The Size and Shape of "Idea Space"

Robin W. Spencer
Director of Research, Imaginatik PLC

published in the
International Journal of Innovation Science (2012) 4:2, 71-76; errata 4:4, 279-280

Abstract

In a large idea management system it is very useful to have a general purpose "ideas like this" capability. Such a tool can be used to define a distance between two ideas, and with a distance metric it is possible to explore the dimensionality and size of a space. Using feature-based Jaccard-Tanimoto similarity, we find that "idea space" is consistently about 14-dimensional regardless of the origin or specifics of the ideas, which has some practical consequences for the behavior and display of similarity search results. In addition, given a distance within which people judge ideas to be "practically identical", the size of the universe of ideas can (whimsically) be estimated at 6 billion ideas.

"IDEAS LIKE THIS"

Idea management systems are increasingly prevalent in large organizations. Internet technology makes it straightforward to solicit and collect ideas from employees and/or customers at very large scale. These systems are especially effective when used to support a challenge-based process in which a specific business need is posed to "the crowd", which opines, comments, and sometimes self-evaluates in an on-line forum, followed by a structured review and decision process which coalesces the input to suit the original sponsor's initiative.

Sustained and successful use of a challenge-based idea management system gives rise to two related observations. First, in very large or fast-moving challenges there is often considerable duplication and overlap of submitted ideas. This poses a burden to reviewers, who typically appoint one or two people to "bucket" the hundreds of submitted ideas into a manageable number of related themes before imposing on the time of subject-matter experts. Secondly, over time a system accumulates thousands of ideas on hundreds of business-relevant topics, of such diversity that traditional exact-query database searching is of limited value. Browsing the content of a large idea management system frequently takes the form of "I don't know exactly what I want, but I'll know it when I see it."

Both of these situations can be improved with a robust "ideas like this" tool, which when presented with any given idea in the system will return a sorted list of the "most similar" entries. We have implemented such a capability using feature counting, where a feature can be a word or phrase with statistically significant frequency, or the author's name, or specific text from popup-menu selections, or if available, background database demographics about the author (his or her department, location, title, etc.). Features of

an idea can also include the results of others' actions, such as codified review results, page hits, comments, and various crowd-rating results. There are many similarity functions based on counting features present in either or both of two objects; we typically use the well-characterized Jaccard or Tanimoto similarity [1].

MEASURING THE DIMENSION OF A SPACE OF OBJECTS

The dimension of any space of objects can be measured simply by counting as a function of distance. For an intuitive example, consider Figure 1 in which pennies are scattered on a table. Pick any penny as the center, and with a simple ruler (with any units), count how many other pennies are within radius **r** of the center penny. Repeat many times to get a sizable table of **n** (the observed number of pennies) as a function of **r**. It is no surprise that for the pennies-on-a-table example, n is found to be proportional to r^2 (because the area of a circle is πr^2). In general, $n = Cr^D$, where C is a constant and D is the dimension of the space. This is the Hausdorff dimension, made familiar by the question that Mandelbrot used to introduce fractal geometry: how long is the coastline of Britain? [2].

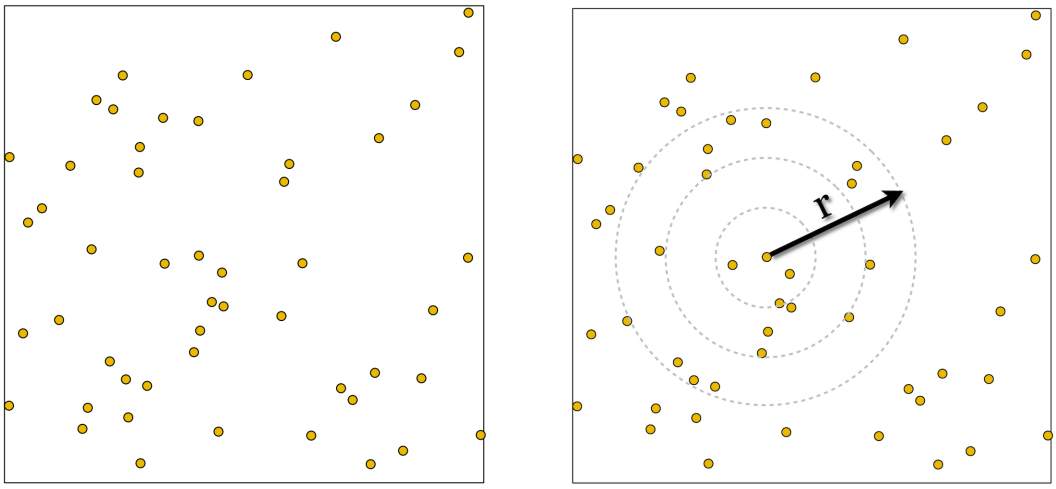

Figure 1. How to find the dimension of a space by counting. (left) Pennies scattered at random on a table. (right) Pick any penny and count how many other pennies n are within radius r of it, for several values of r. The dimension is the slope of a plot of log(**n**) vs log(**r**).

Figure 2 shows results for three representative challenges: one with 185 ideas about energy saving in a large corporate client, another with 470 proposals for spin-off new ventures in a large client, and the third a "wish list" with 620 ideas for new features in our own product line.

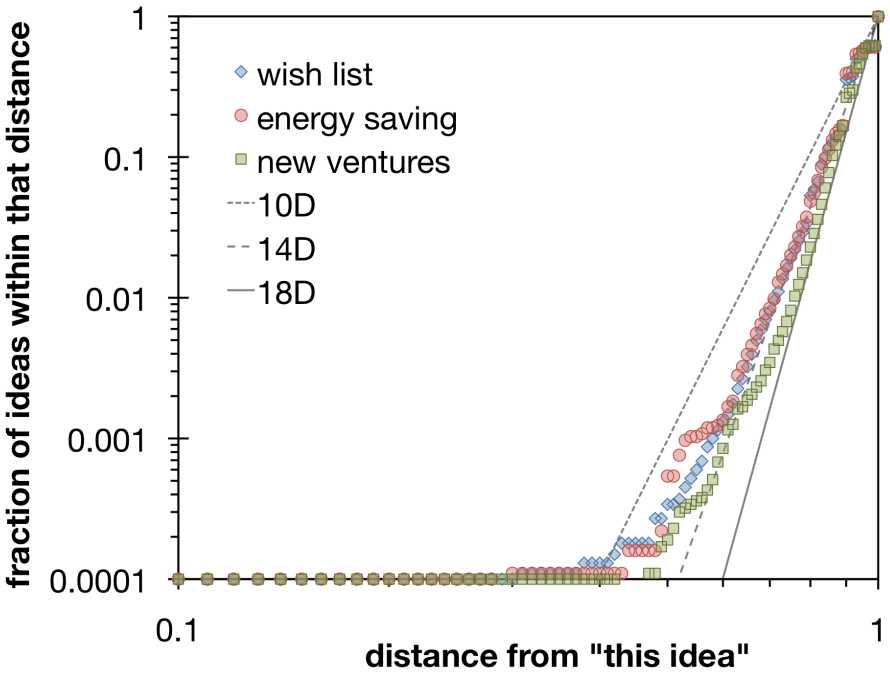

Figure 2. Fraction of ideas within a given distance of a randomly chosen idea for three corporate innovation challenges. Diamonds: 620 suggested features for our own products; circles, 185 ideas for energy saving; squares, 470 spin-off venture proposals. Lines show the behavior of 10, 14, and 18 dimensional spaces (dotted, dashed, solid, respectively). Note the log-log scales.

In each case an idea was selected at random, the Tanimoto feature distance from it to every other idea in the challenge was determined, these distances were binned at every 0.01 unit of distance and the count of how many ideas in at that distance summed, then normalized to the total number of ideas in that dataset. This was repeated 100 times per challenge and the results averaged to produce Figure 2. In such data the total number of ideas is known, and the maximum possible distance is also known (Jaccard or Tanimoto distances must be between 0 and 1). Therefore the dimension equation can be normalized to give

$$n = N(r/R)^D \quad \text{or} \quad \log(n/N) = D \log(r/R)$$

where n = number of ideas within radius r, N = total number of ideas, R = maximum possible radius = 1. It is apparent from Figure 2 that the dimension of idea space as we have defined it is about 14. Repeating the same analysis using Dice distance (= 1-Dice similarity) changes the slopes such that the apparent dimension is about 8. There is little reason to prefer one similarity measure over another [1]; we use Tanimoto because of its extensive use in defining chemical similarity (vide infra). In any case, whether the apparent dimension of idea space is 8 or 14, the point is that it is not 2 or 3, nor is it 100 or more.

HOW CLOSE IS "PRACTICALLY IDENTICAL"? HOW FAR IS "NOT AT ALL RELATED" ?

Given that our distance measure has reassuringly similar behavior for all of the datasets examined, the next task is to use human interviews to set rules-of-thumb that will control how many ideas are suggested to be relevant to the user's current selection.

Note in the left half Figure 2 that there are very few ideas that fall with a distance of 0.2 (= similarity of 0.8) to any other; the curves lift off only above about a distance of 0.4. At the right end of the scale, about 90% of all ideas are distance 0.8 or greater from any given idea (similarity 0.2 or less), that is, not at all similar. We can expect the "interesting stuff" to happen between these limits.

Our experience is that people find ideas with distance greater than 0.75 to 0.8 "too random" to be useful in helping them pursue a train of thought. Distances of 0.6 to 0.75 are relevant and helpful, of 0.2 to 0.6 even more relevant but very rare, and distances less than 0.2 are essentially identical. Fortunately for this purpose our datasets include a handful of duplications (from users copying and pasting), and these are reliably identified within distance 0.2. This highlights a pragmatic reason to seek a threshold for "practically identical": ideas found to be this close are likely to be duplicates or copies, and can be electronically consolidated with little risk of human disagreement.

PACKING A SPACE: HOW MANY IDEAS CAN THERE EVER BE?

We can have fun with the concept of "practically identical" by seeing how many such ideas it takes to fill the universe of idea space.

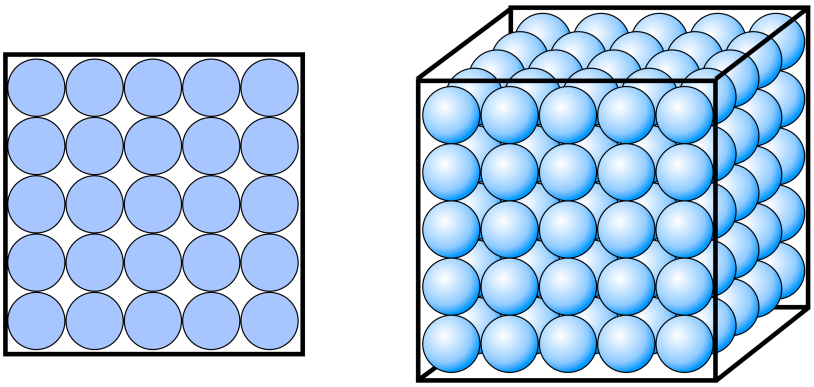

Figure 3. Packing spaces where the maximum distance is 1 and the "identical" distance is 0.2.

Figure 3 shows that if idea space were 2 dimensional, its maximum distance were 1 and the "practically identical" radius were 0.2, that 25 ideas would fill it. That is, any additional idea in the space would have to be within 0.2 of one already existing, and therefore be redundant. For three dimensions, there could be 5 cubed or 125 ideas (Figure 3, right). Clearly the formula for packing is $N = (1/r)^D$, where r is the identity radius and D the dimension of the space. From our estimates, the number of possible non-redundant ideas is $(1/0.2)^{14}$ = 6 billion.

DISCUSSION

The concept of an abstract space with distance, dimension, and size is well established in pharmaceutical research where "chemistry space" is a useful construct based on similarities between molecules. You don't need any chemistry background to see the similarity illustrated in Figure 4.

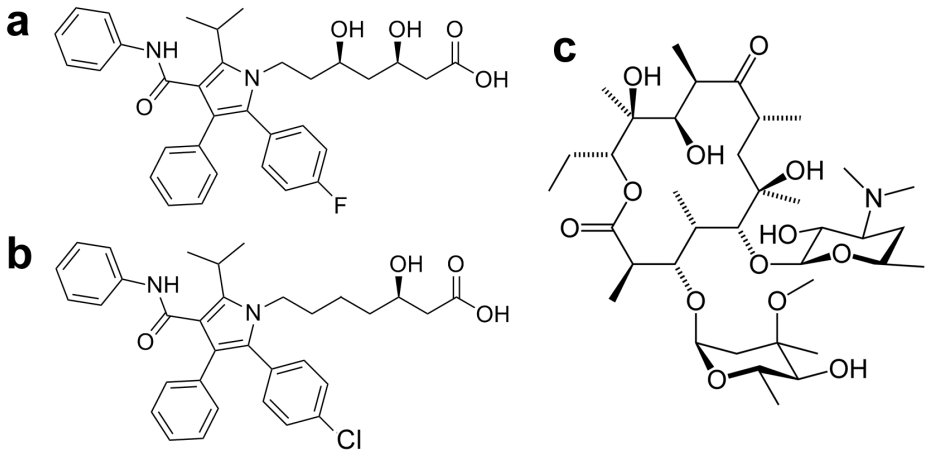

Figure 4. Three molecules: (a) Lipitor, (b) a Lipitor analog, (c) erythromycin.

Lipitor, a popular cholesterol-lowering drug, is shown in Figure 4a, and Figure 4b is lipitor with a hydroxy group removed and chlorine substituted for fluorine. This molecule is quite similar to lipitor and might be expected to have similar biological properties, be made in a similar way, and fall within the scope of lipitor patents. Figure 4c is erythromycin, an important antibiotic. Aside from the fact that it is also an important drug, it has almost nothing in common with lipitor: its different atoms are connected in very different ways. It can be expected to have very different properties.

Feature counting is also used to describe the similarity between molecules. A molecule like those shown might have a fingerprint composed of thousands of features, for example whether or not it contains a benzene ring with a fluorine attached (which is true for lipitor but neither of the other examples). Tanimoto distances are readily computed from such fingerprints and can be used to create a "chemistry space" [3], and the distance within which molecules are likely to have similar biological properties is about 0.15 [4].

The general uses of molecular similarity in drug research are searching, sampling, and bundling. These are listed in Table 1 with analogous uses for a measure of similarity of ideas in innovation and knowledge management.

activity	drug research	idea management
searching	"Molecule X is interesting. Find others like it in our collections or in the patent literature."	"Idea X is interesting. Find others like it in our data systems. Find the people who had similar ideas."
sampling	"Experimental screening is expensive. Find representatives of all clusters of highly similar molecules, so I can test just a subset with confidence that I won't miss a major trend."	"Review experts don't have the time to examine hundreds of ideas. Find representatives of all clusters of highly similar ideas, so I can present them with a subset and still cover all major concepts."
bundling	"Collect all molecules similar to X, so we can find efficiencies in how they're prepared, or we can build a well-exemplified patent estate."	"Collect all ideas similar to X, so we can combine them to make a richer, more detailed proposal."

Table 1: Examples of uses of a similarity function in chemistry and idea management.

Because thousands of simple features are often used in chemistry Tanimoto calculations, chemistry space is sometimes assumed be a thousand-dimensional space [5], though this doesn't take any account of the overlap and redundancy of features, nor the fact that many have negligible effect on physical or biological properties of the molecules they describe. A more formal way to approach the dimensionality of chemistry space would be to take the same large matrix of molecules and their fingerprint vectors, subject it to principal component analysis and keep only the eigenvectors (dimensions) whose eigenvalues are greater than one. This is standard practice in dimensionality reduction but computationally intense and unnecessary for our qualitative purposes. Counting molecules within a given Tanimoto distance of a randomly picked "center molecule" (as in Figure 2) is conceptually simple and gives a Hausdorff dimension of chemistry space of about 7 [6].

There are at least two useful results from our measurement that idea space is about 14 dimensional. Firstly, this value serves as guide to how many different kinds of features it may take to adequately describe the distance between ideas. If we picked just two features (for example author name and author location), we could expect that because $2 << 14$ this would not result in a good measurement: it's equivalent to assuming that all ideas from Alice Jones in London are the same and very different from those of Bob Smith in Chicago. Conversely, we probably do not need many more than 14 types of features because we could expect overlap and redundancy at the cost of collecting and computing these extra data.

The second practical result of estimating the dimensionality of idea space is that it provides a guide to designing intuitive metrics and displays. Abstract spaces are frequently crushed into two dimensions for representation on paper or computer screens, and this is often done with a force-directed algorithm like multidimensional scaling [7]. When three dimensions are squashed into two the results can be quite useful and intuitive especially over small distances; this is our experience with road maps, where at least for distance scales of 1000 miles or less there is little error or confusion in draw-

ing a flat map of our three-dimensional world. But when we know that our idea space is inherently about 14 dimensional, we are forewarned that any representation in two dimensions will either embed considerable error, or we should devise a different distance scale that results in the equivalent of about two dimensions.

REFERENCES

1. Ibrahimov, O., Sethi, I., Dimitrova, N., *Novel Similarity Based Clustering Algorithm for Grouping Broadcast News.* Proc. of SPIE Conf. "Data Mining and Knowledge Discovery: Theory, Tools, and Technology IV", 4730, 394-304, 2002, and Ellis, D., Furner-Hines, J., Willett, P., *Measuring the degree of similarity between objects in text retrieval systems.* Perspectives in Information Management, 3(2), 128-149,1993
2. B. Mandelbrot, *How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension,* Science 156, 636-638, 1967
3. Johnson, A., Maggiora, G. *Concepts and Applications of Molecular Similarity*, Wiley, New York, 1990, and Chen, X., Reynolds, C.H., *Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients*, J Chem Inf Comput Sci. 42(6), 1407-14, 2002
4. Patterson, D. E., Cramer, R. D., Ferguson, A. M., Clark, R. D., Weinberger, L. E. *Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors.* J. Med. Chem. 39, 3049-3059, 1996
5. Pearlman, R., Smith, K., *Novel software tools for chemical diversity*, in Perspectives in Drug Discovery, 9(11), 339-353, 1998
6. Spencer, R., *Can the Size and Shape of Chemistry Space be Biased for More Successful High-Throughput Screening?*, Global Drug Discovery, IBC Library Series, L. M. Savage, ed., 185-218, 1998
7. van der Maaten, L., Postma, E., van den Herik, H. *Dimensionality Reduction: A Comparative Review*, IEEE Transactions on Pattern Analysis and Machine Intelligence, preprint, 2007