How to Tier a Media List with Unsupervised Machine Learning

Michael Burke, MSR Communications

In PR we often face a challenge when our clients or businesses ask us to identify the ‘most relevant’ media outlets. How would you rank the media outlets we should be pursuing based on importance? It’s not an easy question, because no single metric provides a satisfactory answer. Ranking them by circulation is easy enough, but we know all too well that having a high circulation does not mean that an article in the outlet will have a great impact for our brand. We can talk about social media reach, authority and other factors that play into it, and how it’s complicated, but here’s what marketing executives hear: “I don’t know which outlets are the most important.”

Fortunately, this is a problem that’s solvable with a machine learning technique called “clustering”, which can be carried out with software that does not cost anything to download or use.

A crash course in ‘unsupervised learning’

Machine learning is divided into two general types of ‘learning’: supervised, and unsupervised. With supervised learning, you start with lots of examples of something for which you know the answer, which you can feed to the computer program, that allows it to determine the probability of something being similar. For example, to train a computer model to identify cats, you’d feed it a bunch of pictures of different animals, including lots of photos of cats, and let it learn the general characteristics that distinguish cats from dogs, racoons, monkeys and other animals. The key with this approach is that you’d need to already know what a cat is, and be able to provide data that not only gave examples of cats, but also labeled the photos of cats–if you can't tell the program "this is a cat", it doesn't learn anything.

But what if you didn't know what a cat was? Let's say that you had 10,000 photos of furry animals, and you knew they were of different species, but weren't sure how to classify them. "Clustering" could be used to group them according to similarities. Turn a well-tuned clustering model loose on a bunch of digital photos of animals, and it could fairly quickly divide them into buckets according to species. And if you were to examine those groupings, you'd find that for the most part, cats are grouped with cats, dogs with dogs and so forth.

This machine learning technique can also be used to group target media outlets according to similarities that can also tell us the relevance of a certain outlet to our company or client. We don't know the 'outcome', or the answer to the question of how *relevant* the outlet is, but the data we have provides clues that we can use to classify outlets by metrics that indicate relevancy.

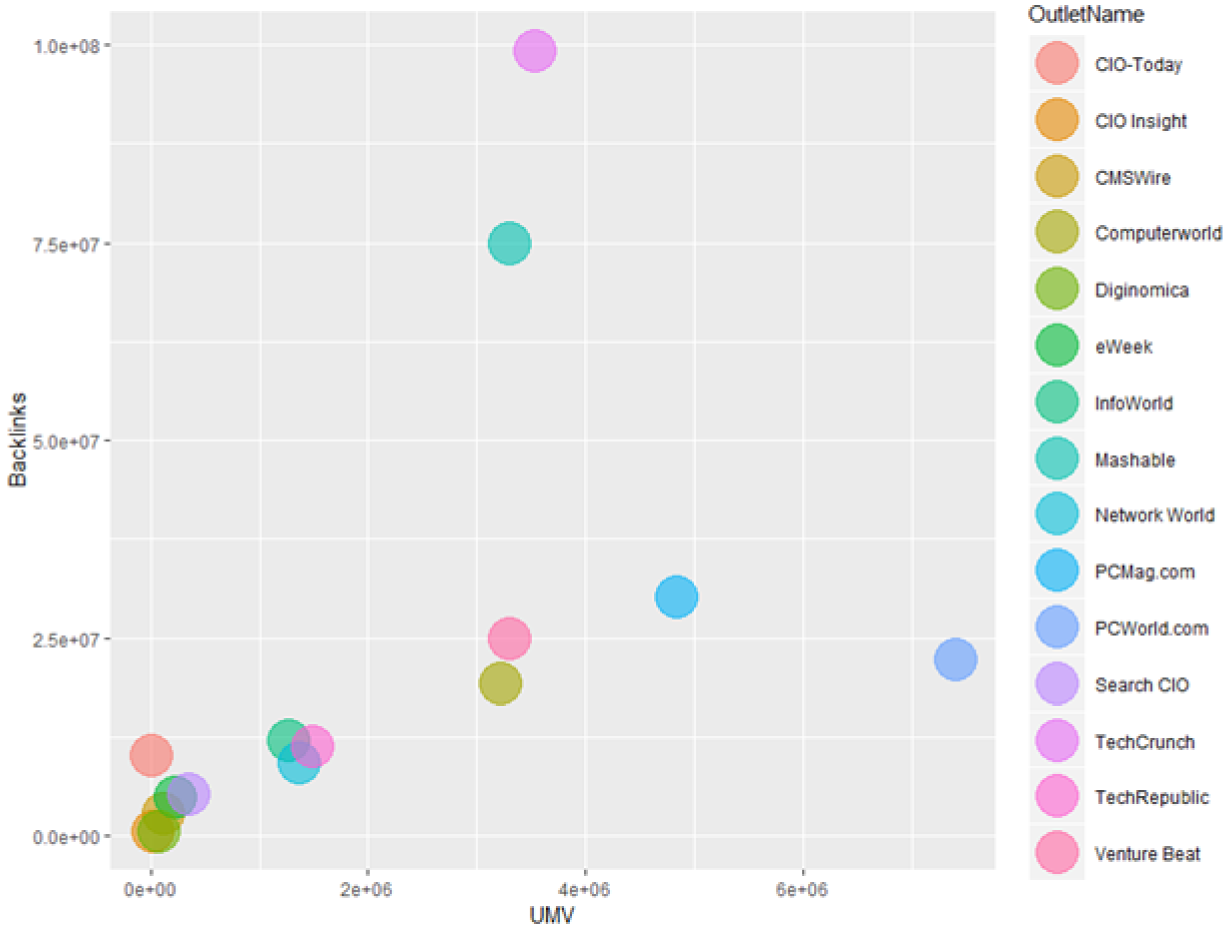
An exercise in clustering: tiering your media list

We had a client that had developed an offering for tech companies, and was very interested in breaking into the tech trade publications. Naturally, they wanted to know which ones would be the most effective venues for the messages. We put gathered a list of 16 top outlets and used SEM Rush to gather information on a variety of metrics that we believe indicate value, including:

- UMV
- # of Backlinks
- Authority Score
- # of top site search terms that were relevant to our

brand

It’s easy to visualize how clustering might work if you look at only *two* metrics. For example, if you look at UMV plotted against # of backlinks, you get something like:



Looking at this chart, you can easily see how there are several different groups of publications that have similar UMVs and backlink volumes. However, UV and backlinks represent only *half* of our data—we also want to know how authority score and the number of relevant search terms works into the mix. We’ve got a bit of a problem here, as when we start throwing in more dimensions, it becomes increasingly difficult for human beings to calculate, or even conceptualize—our brains can’t visualize more than 3 dimensions. A second problem is that all of these metrics are measured on different scales. If we take the numbers at face value, UMVs and backlinks, which have values extending into the millions, will completely overshadow authority score, which has a zero-100 score. The “number of search terms”, which in our case were typically represented by values of no more than six or seven, would be completely overlooked. None of these issues, however, are a problem for a computer, which can easily normalize the data so that it’s analyzed on the same scale, and can create ‘groupings’ based on more than 3 dimensions.

Once we collected our data, we applied a clustering technique known as “hierarchical clustering” to model all four dimensions of the dataset: UMV, Backlinks, Search Terms and Authority Score and rank them according to similarity. We used a tool called RStudio, which can be downloaded and run for free, and is a popular tool with data scientists, as well as scientists in other disciplines. RStudio does require some coding skills, but other tools such as Tableau can perform clustering with zero coding.

Clustering performs its voodoo

Without getting into the math, hierarchical clustering basically calculates a distance between each data point in our dataset, and every other data point in the dataset along multiple dimensions, and then discovers which are closest to each other. Even doing this for two dimensions would be a labor-intensive feat if one were to calculate—doing it across more than two dimensions would be so time-consuming that there’d be no point (which is why almost no one did this before computers).

Applying hierarchical clustering, we discovered five groups of publications in this list that were similar across all four dimensions.

- **Group 1, our top tier targets included:**
    - Computerworld, Network World, Tech Republic and eWeek
- **Group 2, our second tier targets, included:**
    - InfoWorld, CMSWire, SearchCIO CIO-Today

- **Group 3, our third tier targets, included:**
    - Mashable, TechCrunch
- **Group 4, our fourth tier targets, included:**
    - PCMag.com, PCWorld, Venture Beat
- **Group 5, our lowest tier targets, included:**
    - CIO Insight, Diginomica.

As Mashable and TechCrunch tend to be highly coveted outlets, why aren't they in a higher tier? In our case, it is most likely because they have low numbers of relevant search terms. But had you just been looking at the data without the aid of clustering, you wouldn't necessarily know how much those actually affect the relevance, and probably would have assumed that they were top-tier for your client.

In sum, the results that you get from clustering will vary greatly based on the data that you include in the mix. We chose to look at UMV, backlinks, authority score and search terms, but you might decide that other metrics are more useful. For example, you might decide that social media reach is a critical metric, and work this into the model. In this way, this technique can be tailored to your exact needs.

***About the Author:** Michael Burke has worked with some of the world's top brands on marketing and PR strategy, including The Myers-Briggs Company and AirBnB, as well as dozens of cutting edge technology clients. As a director and data scientist at MSR Communications, he's living his dream of applying data science to MarComm and PR.*