Effective Software Design

Doing the right thing.

Data Science: Set Similarity Metrics

Posted on February 27, 2019 by Hayim Makabee

In this article I would like to present several metrics to calculate the similarity between sets of items. I've been analyzing diverse metrics as part of an investigation I'm doing to improve the targeting of digital advertising campaigns. These set similarity metrics are very useful to address the problem of audience expansion.

The basic idea is: we start with a set of users that have engaged with the ad, for example clickers. Then we try to find other similar sets of users that we can target. We have the expectation that, because of their similarity, the targeted users will also engage positively with the ad. These similar users become our expanded audience.

Jaccard Similarity

The Jaccard Similarity is defined as the size of the intersection divided by the size of the union of the sets.

Given two sets, A and B, the Jaccard Similarity is defined as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

The Jaccard Similarity ranges between zero and one.

Also called: Jaccard index, Intersection over Union

For more information: Jaccard Similarity

Sorensen Coefficient

The Sorensen Coefficient equals twice the number of elements common to both sets divided by the sum of the number of elements in each set.

Given two sets, X and Y, the Sorensen Coefficient is defined as:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}$$

The Sorensen Coefficient ranges between zero and one.

Also called: Sorensen–Dice index, Sorensen index, Dice's coefficient

For more information: Sorensen Coefficient

Tversky Index

For sets X and Y, the Tversky Index is given by:

$$S(X,Y) = \frac{|X \cap Y|}{|X \cap Y| + \alpha|X - Y| + \beta|Y - X|}$$

Note that $\alpha, \beta \geq 0$ are parameters of the Tversky Index.

The Tversky Index ranges between zero and one.

The Tversky Index can be seen as a generalization of the Jaccard Similarity and the Sorensen Coefficient:

- Setting $\alpha = \beta = 1$ produces the Jaccard Similarity.
- Setting $\alpha = \beta = 0.5$ produces the Sorensen Coefficient.

Tversky measures with $\alpha + \beta = 1$ are of special interest.

For more information: Tversky Index

Overlap Coefficient

The Overlap Coefficient is defined as the size of the intersection divided by the size of the smaller of the two sets.

For sets X and Y, the Overlap Coefficient is given by:

$$\mathrm{overlap}(X,Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)}$$

If set X is a subset of Y or the converse then the Overlap Coefficient is equal to 1.

Also called: Szymkiewicz–Simpson Coefficient

For more information: Overlap Coefficient

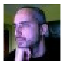
About Hayim Makabee

Veteran software developer, enthusiastic programmer, author of a book on Object-Oriented Programming, co-founder and CEO at KashKlik, an innovative Influencer Marketing platform.

View all posts by Hayim Makabee →

This entry was posted in Data Science, Machine Learning and tagged Data Science, Machine Learning. Bookmark the permalink.

5 Responses to *Data Science: Set Similarity Metrics*



Lior Kogan *says:*

March 3, 2019 at 11:34 am

Nice!

Also Hamming Distance between finite sets:

http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.224.4799 https://stackoverflow.com/questions/29425742/calculate-the-hamming-distance-between-the-two-same-datasets/29428071#29428071

Reply

Hayim Makabee *says:*

March 3, 2019 at 11:47 am

Thanks Lior, very nice!

Reply

vhen *says:*

March 22, 2019 at 8:39 am

Not to be nitpicky, but for the overlap coefficient, when you write:
"The Overlap Coefficient is defined as the size of the intersection divided by the smaller of the size of the two sets."
I think you mean:
"The Overlap Coefficient is defined as the size of the intersection divided by the size of the smaller of the two sets."

Reply

Hayim Makabee *says:*

March 27, 2019 at 11:51 am

Thanks! I fixed it now. You are invited to read and comment on my other blog posts as well. 😉

Reply

Pingback: *String Similarity Metrics: Token Methods | Baeldung on Computer Science*

Effective Software Design

Blog at WordPress.com.