Create account Log in

Article Talk Read Edit View history Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Languages
Español
中文
Edit links

Sørensen–Dice coefficient

From Wikipedia, the free encyclopedia
(Redirected from Dice's coefficient)

The **Sørensen–Dice index**, also known by other names (see Names, below), is a statistic used for comparing the similarity of two samples. It was independently developed by the botanists Thorvald Sørensen[1] and Lee Raymond Dice,[2] who published in 1948 and 1945 respectively.

Contents [hide]

1 Name
2 Quantitative version
3 Formula
4 Difference from Jaccard
5 Applications
6 See also
7 References
8 External links

Name [edit]

The index is known by several other names, usually **Sørensen index** or **Dice's coefficient**. Both names also see "similarity coefficient", "index", and other such variations. Common alternate spellings for Sørensen are Sorenson, Soerenson index and Sörenson index, and all three can also be seen with the –sen ending.

Other names include:

- Czekanowski's binary (non-quantitative) index[3]

Quantitative version [edit]

The expression is easily extended to abundance instead of presence/absence of species. This quantitative version is known by several names:

- Quantitative Sørensen–Dice index[3]
- Quantitative Sørensen index[3]
- Quantitative Dice index[3]
- Bray-Curtis similarity (1 minus the *Bray-Curtis dissimilarity*)[3]
- Czekanowski's quantitative index[3]
- Steinhaus index[3]
- Pielou's percentage similarity[3]
- 1 minus the Hellinger distance[4]

Formula [edit]

Sørensen's original formula was intended to be applied to presence/absence data, and is

$$QS = \frac{2C}{A+B} = \frac{2|A \cap B|}{|A|+|B|}$$

where *A* and *B* are the number of species in samples A and B, respectively, and *C* is the number of species shared by the two samples; QS is the quotient of similarity and ranges between 0 and 1.[5]

It can be viewed as a similarity measure over sets:

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

Similarly to Jaccard, the set operations can be expressed in terms of vector operations over binary vectors *A* and *B*:

$$s_v = \frac{2|A \cdot B|}{|A|^2 + |B|^2}$$

which gives the same outcome over binary vectors and also gives a more general similarity metric over vectors in general terms.

For sets *X* and *Y* of keywords used in information retrieval, the coefficient may be defined as twice the shared information (intersection) over the sum of cardinalities :[6]

When taken as a string similarity measure, the coefficient may be calculated for two strings, *x* and *y* using bigrams as follows:[7]

$$s = \frac{2n_t}{n_x + n_y}$$

where n_t is the number of character bigrams found in both strings, n_x is the number of bigrams in string *x* and n_y is the number of bigrams in string *y*. For example, to calculate the similarity between:

`night`
`nacht`

We would find the set of bigrams in each word:

{`ni`,`ig`,`gh`,`ht`}
{`na`,`ac`,`ch`,`ht`}

Each set has four elements, and the intersection of these two sets has only one element: `ht`.

Inserting these numbers into the formula, we calculate, *s* = (2 · 1) / (4 + 4) = 0.25.

Difference from Jaccard [edit]

This coefficient is not very different in form from the Jaccard index. However, since it doesn't satisfy the triangle inequality, it can be considered a semimetric version of the Jaccard index.[3]

The function ranges between zero and one, like Jaccard. Unlike Jaccard, the corresponding difference function

$$d = 1 - \frac{2|X \cap Y|}{|X| + |Y|}$$

is not a proper distance metric as it does not possess the property of triangle inequality.[3] The simplest counterexample of this is given by the three sets {a}, {b}, and {a,b}, the distance between the first two being 1, and the difference between the third and each of the others being one-third. To satisfy the triangle inequality, the sum of *any* two of these three sides must be greater than or equal to the remaining side. However, the distance between {a} and {a,b} plus the distance between {b} and {a,b} equals 2/3 and is therefore less than the distance between {a} and {b} which is 1.

Applications [edit]

The Sørensen–Dice coefficient is mainly useful for ecological community data (e.g. Looman & Campbell, 1960[8]). Justification for its use is primarily empirical rather than theoretical (although it can be justified theoretically as the intersection of two fuzzy sets[9]). As compared to Euclidean distance, Sørensen distance retains sensitivity in more heterogeneous data sets and gives less weight to outliers.[10] Recently the Dice score (and its variations, e.g. logDice taking a logarithm of it) has become popular in computer lexicography for measuring the lexical association score of two given words.[11]

See also [edit]

- Correlation
- Czekanowski similarity index
- Jaccard index
- Hamming distance
- Horn’s index
- Hurlbert’s index
- Kulczyński similarity index
- Pianka's index
- MacArthur and Levin's index
- Mantel test

- Morisita's overlap index
- Most frequent k characters
- Overlap coefficient
- Renkonen similarity index (due to Olavi Renkonen)
- Simplified Morisita’s index
- Tversky index
- Universal adaptive strategy theory (UAST)

References [edit]

1. ^ Sørensen, T. (1948). "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons". *Kongelige Danske Videnskabernes Selskab* **5** (4): 1–34.
2. ^ Dice, Lee R. (1945). "Measures of the Amount of Ecologic Association Between Species". *Ecology* **26** (3): 297–302. doi:10.2307/1932409. JSTOR 1932409.
3. ^ a b c d e f g h i j Gallagher, E.D., 1999. COMPAH Documentation, University of Massachusetts, Boston
4. ^ Bray, J. Roger; Curtis, J. T. (1957). "An Ordination of the Upland Forest Communities of Southern Wisconsin". *Ecological Monographs* **27** (4): 326–349. doi:10.2307/1942268.
5. ^ http://www.sekj.org/PDF/anbf40/anbf40-415.pdf
6. ^ van Rijsbergen, Cornelis Joost (1979). *Information Retrieval*. London: Butterworths. ISBN 3-642-12274-4.
7. ^ Kondrak, Grzegorz; Marcu, Daniel; Knight, Kevin (2003). "Cognates Can Improve Statistical Translation Models" (PDF). *Proceedings of HLT-NAACL 2003: Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. pp. 46–48.
8. ^ Looman, J. and Campbell, J.B. (1960) Adaptation of Sorensen's K (1948) for estimating unit affinities in prairie vegetation. Ecology 41 (3): 409–416.
9. ^ Roberts, D.W. (1986) Ordination on the basis of fuzzy set theory. Vegetatio 66 (3): 123–131.
10. ^ McCune, Bruce & Grace, James (2002) Analysis of Ecological Communities. Mjm Software Design; ISBN 0-9721290-0-6.
11. ^ Rychlý, P. (2008) A lexicographer-friendly association score. Proceedings of the Second Workshop on Recent Advances in Slavonic Natural Language Processing RASLAN 2008: 6–9

External links [edit]

- Open Source Dice / Sorensen Scala implementation as part of the larger stringmetric project

The Wikibook *Algorithm implementation* has a page on the topic of: ***Dice's coefficient***

Categories: Information retrieval evaluation | String similarity measures | Measure theory

This page was last modified on 13 June 2015, at 03:36.