An Information-Theoretic Definition of Similarity

Dingquan Li

Peking University

dingquanli@pku.edu.cn

December 3, 2018

Paper Information

Dekang Lin, An information-theoretic definition of similarity, ICML, 1998.

- Co-founder, CTO@Naturali
- Senior Staff Research Scientist@Google
- Professor@University of Alberta

Contributions

- An information-theoretic definition of similarity that is applicable as long as there is a probabilistic model
- Demonstrate how the definition can be used to measure the similarity in a number of different domains

Overview

Overview

Problems

- Each of the previous similarity measures are tied to a particular application or assume a particular domain model.
- Their underlying assumptions are often not explicitly stated. Almost all of the comparisons and evaluations are based on empirical results.

Solutions

- **Universality**: Define similarity in information-theoretic terms, which is applicable as long as the domain has a probabilistic model.
- **Theoretical Justification**: The similarity measure is derived from a set of assumptions about similarity. If the assumptions are deemed reasonable, the similarity measure necessarily follows.

Overview

Intuitions

1. The similarity between A and B is related to their commonality. The more commonality they share, the more similar they are.
2. The similarity between A and B is related to the differences between them. The more differences they have, the less similar they are.
3. The maximum similarity between A and B is reached when they are identical, no matter how much commonality they share.

Assumption 1: The commonality between A and B is measured by

$$I(\mathrm{common}(A, B)),$$

where $\mathrm{common}(A, B)$ is a proposition that states the commonalities between A and B; $I(s)$ is the amount of information contained in a proposition s.
In information theory, the information contained in a statement is measured by the negative logarithm of the probability of the statement. Therefore,

$$I(\mathrm{common}(A, B)) = -\log P(\mathrm{common}(A, B)).$$

Assumptions

Assumption 2: The differences between A and B is measured by

$$I(\mathrm{description}(A, B)) - I(\mathrm{common}(A, B)),$$

where $\mathrm{description}(A, B)$ is a proposition that describes what A and B are.

Assumptions

Assumption 3: The similarity between A and B, $\mathrm{sim}(A, B)$, is a function of their commonalities and differences. That is,

$$\mathrm{sim}(A, B) = f(I(\mathrm{common}(A, B)), I(\mathrm{description}(A, B))),$$

where the domain of f is $\{(x, y)|x \geq 0, y > 0, y \geq x\}$.

Assumptions

Assumption 4: The similarity between a pair of identical objects is 1. When A and B are identical, knowing their commonalities means knowing what they are, *i.e.*, $I(\text{common}(A, B)) = I(\text{description}(A, B))$. Therefore, the function f must have the property: $\forall x > 0, f(x, x) = 1$.

Assumptions

Assumption 5: When there is no commonality between A and B, their similarity is 0, no matter how different they are.

$$\forall y > 0, f(0, y) = 0.$$

Assumptions

Assumption 6: The overall similarity of the two objects is a weighted average of their similarities computed from different perspectives.

$$\forall x_1 \leq y_1, x_2 \leq y_2, f(x_1 + x_2, y_1 + y_2) = \frac{y_1}{y_1 + y_2} f(x_1, y_1) + \frac{y_2}{y_1 + y_2} f(x_2, y_2).$$

Similarity Theorem

Theorem (Similarity Theorem)

Under the above six assumptions, the similarity between A and B is measured by the ratio between the amount of information needed to state the commonality of A and B and the information needed to fully describe what A and B are:

$$\mathrm{sim}(A, B) = \frac{\log P(\mathrm{common}(A, B))}{\log P(\mathrm{description}(A, B))}$$

Similarity Theorem

Proof.

For $y = x$, we have $f(x, y) = f(x, x) = 1 = \frac{x}{y}$.
For $y > x$, based on Assumptions 4,5,6, we have

$$f(x, y) = f(x + 0, x + (y - x)) = \frac{x}{y} f(x, x) + \frac{y - x}{y} f(0, y - x)$$
$$= \frac{x}{y} \cdot 1 + \frac{y - x}{y} \cdot 0 = \frac{x}{y}$$

$\square$

Similarity Theorem

Note: If we know the commonality of the two objects, their similarity tells us how much more information is needed to determine what these two objects are.

Overview

Similarity between Ordinal Values

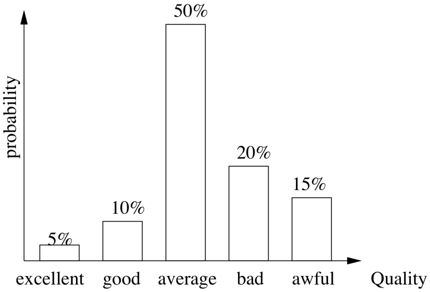


Figure: Example Distribution of Ordinal Values

$$\text{sim}(\text{excellent}, \text{good}) = \frac{\log P^2(\text{excellent} \vee \text{good})}{\log P(\text{excellent})P(\text{good})} = 0.72$$

Similarity between Ordinal Values

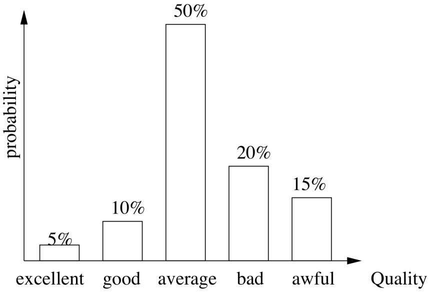


Figure: Example Distribution of Ordinal Values

$$\text{sim}(\text{good}, \text{average}) = \frac{\log P^2(\text{good} \vee \text{average})}{\log P(\text{good})P(\text{average})} = 0.34$$

Similarity between Ordinal Values

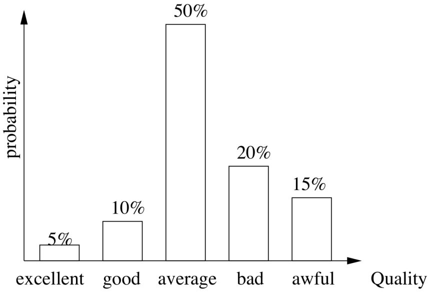


Figure: Example Distribution of Ordinal Values

$$\text{sim}(\text{excellent}, \text{average}) = \frac{\log P^2(\text{excellent} \vee \text{good} \vee \text{average})}{\log P(\text{excellent}) P(\text{average})} = 0.23$$

Similarity between Ordinal Values

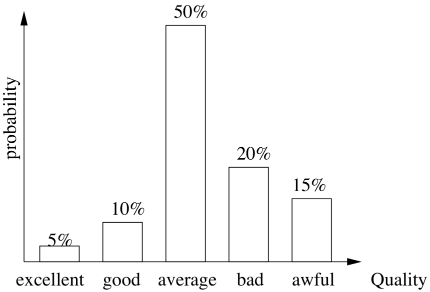


Figure: Example Distribution of Ordinal Values

$$\mathrm{sim}(\mathrm{good}, \mathrm{bad}) = \frac{\log P^2(\mathrm{good} \vee \mathrm{average} \vee \mathrm{bad})}{\log P(\mathrm{good})P(\mathrm{bad})} = 0.11$$

Overview

String Similarity — A Case Study

$$\mathrm{sim}_{\mathrm{edit}}(x, y) = \frac{1}{1 + \mathrm{editDist}(x, y)}$$

$$\mathrm{sim}_{\mathrm{tri}}(x, y) = \frac{1}{1 + |\mathrm{tri}(x)| + |\mathrm{tri}(y)| - 2 \star |\mathrm{tri}(x) \cap \mathrm{tri}(y)|}$$

$$\mathrm{sim}(x, y) = \frac{2 \times \sum_{t \in \mathrm{tri}(x) \cap \mathrm{tri}(y)} \log P(t)}{\sum_{t \in \mathrm{tri}(x)} \log P(t) + \sum_{t \in \mathrm{tri}(y)} \log P(t)}$$

String Similarity — A Case Study

Table 1: Top-10 Most Similar Words to "grandiloquent"

Rank	sim_{edit}		sim_{tri}		sim	
1	grandiloquently	1/3	grandiloquently	1/2	grandiloquently	0.92
2	grandiloquence	1/4	grandiloquence	1/4	grandiloquence	0.89
3	magniloquent	1/6	eloquent	1/8	eloquent	0.61
4	gradient	1/6	grand	1/9	magniloquent	0.59
5	grandaunt	1/7	grande	1/10	ineloquent	0.55
6	gradients	1/7	rand	1/10	eloquently	0.55
7	grandiose	1/7	magniloquent	1/10	ineloquently	0.50
8	diluent	1/7	ineloquent	1/10	magniloquence	0.50
9	ineloquent	1/8	grands	1/10	eloquence	0.50
10	grandson	1/8	eloquently	1/10	ventriloquy	0.42

String Similarity — A Case Study

Let W denote the set of words in the word list and W_{root} denote the subset of W that are derived from the same *root* as the given word w (excluding w). Let $(w_1, \cdots, w_n)$ denote the ordering of $W - \{w\}$ in descending similarity to w according to a similarity measure. The precision of $(w_1, \cdots, w_n)$ at recall level $N\%$ is defined as

$$\max_k \quad \frac{|W_{root} \cap \{w_1, \cdots, w_k\}|}{k},$$

$$s.t., \quad \frac{|W_{root} \cap \{w_1, \cdots, w_k\}|}{|W_{root}|} \geq N\%.$$

The quality of $(w_1, \cdots, w_n)$ can be measured by the 11-point average of its precisions at recall levels $0\%, 10\%, 20\%, \cdots, \text{and} 100\%$. The average precision values are then averaged over all the words in W_{root}

String Similarity — A Case Study

Table 2: Evaluation of String Similarity Measures

Root	Meaning	$\|W_{root}\|$	11-point average precisions		
			sim_{edit}	sim_{tri}	sim
agog	leader, leading, bring	23	37%	40%	70%
cardi	heart	56	18%	21%	47%
circum	around, surrounding	58	24%	19%	68%
gress	to step, to walk, to go	84	22%	31%	52%
loqu	to speak	39	19%	20%	57%

Overview

Dependency Triples

Table 3: Features of "duty" and "sanction"

Feature	duty	sanction	$I(f_i)$
f_1: subj-of(include)	x	x	3.15
f_2: obj-of(assume)	x		5.43
f_3: obj-of(avert)	x	x	5.88
f_4: obj-of(ease)		x	4.99
f_5: obj-of(impose)	x	x	4.97
f_6: adj-mod(fiduciary)	x		7.76
f_7: adj-mod(punitive)	x	x	7.10
f_8: adj-mod(economic)		x	3.70

Word Similarity

Let $F(w)$ be the set of features possessed by w. $F(w)$ can be viewed as a description of the word w. The commonalities between two words w_1 and w_2 is then $F(w_1) \cap F(w_2)$.
The similarity between two words is defined as follows:

$$\mathrm{sim} = \frac{2 \times I(F(w_1) \cap F(w_2))}{I(F(w_1)) + I(F(w_2))},$$

where $I(S)$ is the amount of information contained in a set of features S.
Assuming the features are independent of one another,
$I(S) = -\sum_{f \in S} log(P(f))$, where $P(f)$ is the probability of feature f.

Example: "duty"

duty n. 1. obligation, responsibility; onus; business, province. 2. function, task, assignment, charge. 3. tax, tariff, customs, excise, levy.

Respective Nearest Neighbors

Two words are a pair of respective nearest neighbors (RNNs) if each is the others most similar word.

Table 4: Respective Nearest Neighbors

Rank	RNN	Sim
1	earnings profit	0.50
11	revenue sale	0.39
21	acquisition merger	0.34
31	attorney lawyer	0.32
41	data information	0.30
51	amount number	0.27
61	downturn slump	0.26
71	there way	0.24
81	fear worry	0.23
91	jacket shirt	0.22
101	film movie	0.21
111	felony misdemeanor	0.21
121	importance significance	0.20
131	reaction response	0.19
141	heroin marijuana	0.19
151	championship tournament	0.18
161	consequence implication	0.18
171	rape robbery	0.17
181	dinner lunch	0.17
191	turmoil upheaval	0.17
201	biggest largest	0.17
211	blaze fire	0.16
221	captive westerner	0.16
231	imprisonment probation	0.16

Overview

Semantic Similarity in a Taxonomy

The semantic similarity between two classes C_1 and C_2 is not about the classes themselves. $\mathrm{sim}(C_1, C_2)$ is the similarity between x_1 and x_2 if all we know about x_1 and x_2 is that $x_1 \in C_1$ and $x_2 \in C_2$. Assuming that the taxonomy is a tree, if $x_1 \in C_1$ and $x_2 \in C_2$, the commonality between x_1 and x_2 is $x_1 \in C_0 \wedge x_2 \in C_0$, where C_0 is the most specific class that subsumes both C_1 and C_2.

$$\mathrm{sim}(x_1, x_2) = \frac{2 \times \log P(C_0)}{\log P(C_1) + \log P(C_2)}$$

Example

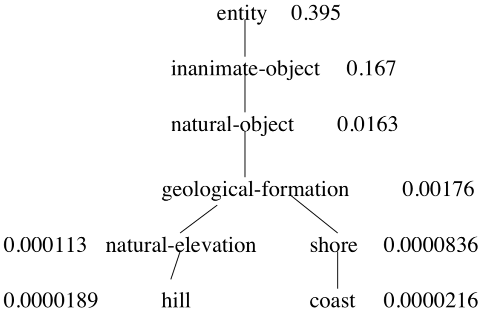


Figure 2: A Fragment of WordNet

$$\mathrm{sim}(\text{hill}, \text{coast}) = \frac{2 \times \log P(\text{geological-formation})}{\log P(\text{hill}) + \log P(\text{coast})} = 0.59$$

Quantitative Results

Word Pair	Miller& Charles	Resnik	Wu & Palmer	sim
car, automobile	3.92	11.630	1.00	1.00
gem, jewel	3.84	15.634	1.00	1.00
journey, voyage	3.84	11.806	.91	.89
boy, lad	3.76	7.003	.90	.85
coast, shore	3.70	9.375	.90	.93
asylum, madhouse	3.61	13.517	.93	.97
magician, wizard	3.50	8.744	1.00	1.00
midday, noon	3.42	11.773	1.00	1.00
furnace, stove	3.11	2.246	.41	.18
food, fruit	3.08	1.703	.33	.24
bird, cock	3.05	8.202	.91	.83
bird, crane	2.97	8.202	.78	.67
tool, implement	2.95	6.136	.90	.80
brother, monk	2.82	1.722	.50	.16
crane, implement	1.68	3.263	.63	.39
lad, brother	1.66	1.722	.55	.20
journey, car	1.16	0	0	0
monk, oracle	1.10	1.722	.41	.14
food, rooster	0.89	.538	.7	.04
coast, hill	0.87	6.329	.63	.58
forest, graveyard	0.84	0	0	0
monk, slave	0.55	1.722	.55	.18
coast, forest	0.42	1.703	.33	.16
lad, wizard	0.42	1.722	.55	.20
chord, smile	0.13	2.947	.41	.20
glass, magician	0.11	.538	.11	.06
noon, string	0.08	0	0	0
rooster, voyage	0.08	0	0	0
Correlation with Miller & Charles	1.00	0.795	0.803	0.834

Overview

Similarity Measures

- **Dice coefficient**

$$\mathrm{sim}_{\mathrm{dice}}(A, B) = \frac{2 \times \sum_{i=1}^{n} a_i b_i}{\sum_{i=1}^{n} a_i^2 + \sum_{i=1}^{n} b_i^2}$$

- **distance-based similarity**

$$\mathrm{sim}_{\mathrm{dist}}(A, B) = \frac{1}{1 + \mathrm{dist}(A, B)}$$

- **Resnik (IJCAI 1995)**

$$\mathrm{sim}_{\mathrm{Resnik}}(A, B) = \frac{1}{2} I(\mathrm{common}(A, B))$$

- **Wu & Palmer (ACL 1994)**

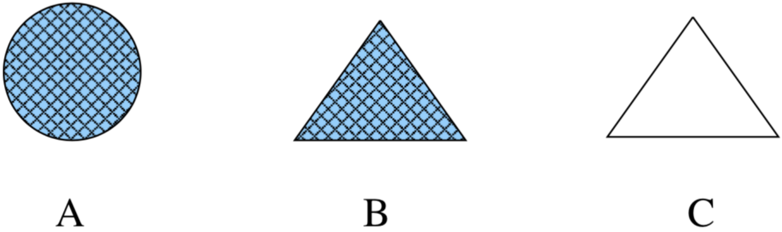
$$\mathrm{sim}_{\mathrm{Wu\&Palmer}}(A, B) = \frac{2 \times N_{CR}}{N_{AC} + N_{BC} + 2 \times N_{CR}}$$

Comparison between Different Similarity Measures

Similarity Measures:
WP: $\text{sim}_{\text{Wu\&Palmer}}$
R: $\text{sim}_{\text{Resnik}}$
Dice: sim_{dice}

Property	sim	WP	R	Dice	sim_{dist}
increase with commonality	yes	yes	yes	yes	no
decrease with difference	yes	yes	no	yes	yes
triangle inequality	no	no	no	no	yes
Assumption 6	yes	yes	no	yes	no
max value=1	yes	yes	no	yes	yes
semantic similarity	yes	yes	yes	no	yes
word similarity	yes	no	no	yes	yes
ordinal values	yes	no	no	no	no

Counter-example of Triangle Inequality



Overview

Conclusion

- A universal definition of similarity in terms of information theory, derived from a set of assumptions.
- The universality of the definition is demonstrated by its applications in different domains

References

Amos Tversky (1977)
Features of similarity.
Psychological Review 84(4), pp. 327 – 352.

Philip Resnik (1995)
Using information content to evaluate semantic similarity in a taxonomy.
IJCAI 1995, pp. 448 – 453.

George A. Miller and Walter G. Charles (1991)
Contextual correlates of semantic similarity.
Language and Cognitive Processes 6(1), pp. 1 – 28.