How Intuitive Is It? Comparing Metrics for Attitudes in Argumentation with a Human Baseline*

Markus Brenneis and Martin Mauve

Heinrich-Heine-Universität, Universitätsstraße 1, 40225 Düsseldorf, Germany
Markus.Brenneis@uni-duesseldorf.de

Abstract. It is often interesting to know how similar two persons argue, e.g. when comparing the attitudes of voters and political parties, or when building an argumentation-based recommender system. Those applications need a distance function, which should give intuitive results. In this paper, we present seven functions which calculate how similar the attitudes of two agents are in an argumentation. We evaluate how good those functions match the results of a human baseline which we determined in a previous work. As it turns out, variants of the p-metric, Cosine, and Soergel distance best agree with human intuition.

Keywords: Argumentation · Metric · Human Baseline.

1 Introduction

Comparing the attitudes different people or organizations have in an argumentation is often relevant and useful, e.g. for clustering using opinions mentioned in argumentations, recommender systems for argumentation platforms (as used in our platform *deliberate* [4]), or comparing one's own attitudes and arguments with those of political parties. In a previous work [5], we have conducted a survey with untrained human subjects to find out what properties a distance function for argumentation data should fulfill to yield results matching human intuition.

In this paper, we compare different distance functions regarding those properties. Our goal is to provide hints for application developers which kinds of distance functions best match human intuition and where and why there are differences. This helps with choosing functions best suited for the problem at hand, knowing that their results follow intuitive and understandable properties.

Our contribution is the following: We present a list of properties which should be fulfilled by a distance function which compares argumentations, based on a survey we have conducted earlier. Different existing distance functions were

* Manchot research group *Decision-making with the help of Artificial Intelligence*, use case politics

The final authenticated version is available online at https://doi.org/10.1007/978-3-030-77772-2_9.

adapted to use them with attitudes in arguments. We compare those functions regarding different properties we found to be intuitive through our survey, and examine different values for the hyperparameters of each function. Afterwards, we explain why certain functions perform better than others.

In the next section, we provide the key definitions used throughout our work. Afterwards, we define the formal mathematical model and distance functions we compared. We then present and discuss the results of comparing the functions with a human baseline, and finally have a look at related work.

2 Definitions

The argumentative terms we use in this paper are based on the IBIS model [10] for argumentation. An argumentation consists of *arguments*, and each argument is formed by two *statements*: a *premise* and a *conclusion*. We call the set of all statements S and the set of all arguments $A \subset S^2$.

A special "statement" is the *issue* I, which denotes the topic of an argumentation and has no conclusions. All premises for arguments with I as the conclusion are referred to as *positions*, and are typically actionable items like "We should build more wind power plants." $P \subset S$ is the set of all positions.

Different persons can have individual views in an argumentation: They can (strongly) agree (denoted as `(+)` (agree), or `+` (strongly agree), respectively) or disagree (`(-)`, `-`) with statements[1], be neutral (`0`) about a statement, indicate to not have an opinion ($\emptyset$), or do not mention anything about a statement (`?`; so we do not know their opinion); we call this stance on statements *opinion*. We define the set of possible opinion values for a statement $O \coloneqq \{\texttt{+}, \texttt{-}, \texttt{(+)}, \texttt{(-)}, \texttt{0}, \emptyset, \texttt{?}\}$.

They can also assign arguments different *relevances* (or *weights* or *importances*), and give a priority order for positions. The overall importances and opinions of a person are referred to as their *attitude*.

We represent a person's attitude as an argumentation tree[2], or, if only positions are involved, as sorted lists with positions, where the most important position is at the top. Note that in our tree representation, statements are nodes, argument are edges, to have statements as atomic building blocks. This visualization can, however, be transformed to classical Dung-based [7] abstract argumentation frameworks when needed. We do not draw the common root I in our visualizations to make them simpler.

As an example, we explain how Alice's tree in Figure 1e should be understood: Alice agrees with the position p and the statements a and b, which build arguments with conclusion p. The argument (a, p) is more important for her than the argument (b, p) (indicated by the bolder edge). Note that we do not differentiate whether an argument edge is attacking or defending – this is up to the interpretation of the natural language presentation of the scenario, but

[1] A more fine-grained model for the strength of (dis-)agreement, as we have suggested in [3], could be used, but is not necessary in this work.

[2] A representation as more general graphs is also possible, but again not necessary for the examples in this work.

is consistent within all trees of one scenario (i.e. in Figure 1e, the edges (a, p) in all three trees are either consistently attacking or supporting arguments); a differentiation is therefore not needed in the model for the purpose of this paper.

Throughout this paper, we use the term *distance function* to refer to a function which calculates some distance between pairs of argumentations with the parts introduced above. Those functions might happen to fulfill all properties of a metric (e.g. the triangle equality), but are not required to do so.

We now define how the drawing of a tree is translated to mathematical objects. Each tree can be considered as a pair of functions (o, s), where $o : S \to O$ captures the opinion on statements, $s : A \to \mathbb{N}_0$ the sorting of arguments by importance (where 1 means top-priority, 0 no priority (as default for not mentioned arguments); the ordering is not required to be injective). Note that we view a function as a set of ordered pairs (parameter, function value).

Please note the following conventions: The sorting position of a position p is treated as the sort order position of a pseudo-argument (p, I). If o is undefined for a value, the function's value is `?`. If s is undefined for a value, the function's value is 0. To keep the notation simple, we assume that the functions' domains are the same when two trees are compared.

For example, Alice's tree in Figure 1e translates to $o = \{(p, \texttt{+}), (a, \texttt{+}), (b, \texttt{+})\}$, $s = \{((a, I), 1), ((a, p), 1), ((b, p), 2)\}$.

A distance function must map the different values to numeric values for calculations. We will evaluate different transformation strategies. As all distance functions need to map the opinion values of O to numeric values and some of them map importance weights to other numeric values, we define the following common mapping strategies:

$$r(x) = \begin{cases} 0.5 & \text{if } x = \texttt{+} \\ 0.25 & \text{if } x = \texttt{(+)} \\ 0 & \text{if } x \in \{\texttt{0}, \texttt{∅}, \texttt{?}\} \\ -0.25 & \text{if } x = \texttt{(-)} \\ -0.5 & \text{if } x = \texttt{-} \end{cases} \tag{1}$$

$$w_h(x) = \frac{1}{x} \tag{2}$$

$$w_g(x) = \frac{1}{2^x} \tag{3}$$

The result of a division by 0 is defined as 0, which means that arguments without importance value (which default to 0) get a calculated weight of 0. The variants $w_{\bar{h}}$ and $w_{\bar{g}}$ are defined the same way, but the values are normalized such that the sum of function values for all arguments with the same conclusion is 1 (or 0, if no argument has a value greater 0). For instance, if we take Alice's tree in Figure 1e again, $w_{\bar{h}}((a, p)) = \frac{\frac{1}{1}}{\frac{1}{1}+\frac{1}{2}} = \frac{2}{3}$. If we mention the function name w, any possible variant can be used (thus, the concrete choice of w is a hyperparameter of the distance function).

Sometimes, we refer to the "simple" opinion, which removes the weight part of the opinion:

$$\text{simple} : O \to \{\texttt{+}, \texttt{-}, \texttt{0}, \emptyset, \texttt{?}\} : x \mapsto \begin{cases} \texttt{+} & \text{if } x = \texttt{(+)} \\ \texttt{-} & \text{if } x = \texttt{(-)} \\ x & \text{otherwise} \end{cases} \tag{4}$$

3 Distance Functions for Argumentations

We now present the distance functions we have compared. Most functions are based on previous work in argumentation theory or related fields and have been adapted by us for use with the formal definition introduced in Section 2. Most functions have hyperparameters, e.g. which function w is used. An overview of the distance functions, their hyperparameters and tested ranges can be found in Table 1.

Table 1: Overview of examined distance functions and their hyperparameters with tested values

Function	Hyperparameters
Bhavsar	$w \in \{w_{\bar{h}}, w_{\bar{g}}\}$, $N \in \{.1, .25, .5, .75, .9\}$
Cosine	$w \in \{w_g, w_{\bar{g}}, w_h, w_{\bar{h}}\}$
Jaccard	$\text{set} \in \{\text{set}_a, \text{set}_o, \text{set}_s, \text{set}_{s'}\}$, $\text{keep} \in \{\text{keep}_a, \text{keep}_t\}$
p-metric	$p \in \{1, 2\}$, $ds \in \{d_{s_w}, d_{s_s}\}$, $da \in \{da_0, da_s\}$, $w \in \{w_g, w_{\bar{g}}, w_h, w_{\bar{h}}\}$
Soergel	$w \in \{w_g, w_{\bar{g}}, w_h, w_{\bar{h}}\}$
VAA	–
WATD	$\alpha \in \{.1, .25, .5, .75, .9\}$, $w \in \{w_g, w_{\bar{g}}, w_h, w_{\bar{h}}\}$

Bhavsar distance [2] presented a metric for match-making of agents in e-business environments, which are represented as trees. As the definition of that recursive metric is lengthy, we do not repeat its definition here. The metric can be applied to our structure by transforming sort orders using $w_{\bar{h}}$ or $w_{\bar{g}}$, and treating opinions as node labels. A parameter N sets the relative importance of subtrees and respective roots, similar to the PageRank algorithm [15].

Cosine distance We define the Cosine distance similar to [16], who predict opinions in argumentation. They treat accepting and declining a statement s as two different entities ("acceptance of s" and "acceptance of $\neg s$") and ignore a statement if it has no rating in one of the inputs:

$$d(t_1, t_2) = 1 - \frac{V_1 \cdot V_2}{||V_1||\,||V_2||} \tag{5}$$

where an argumentation tree $t_i = (o_i, s_i)$ is transformed to a vector V_i with the components $s_i(a)$ for every argument a, and $\max(-r(o_i(s)), 0)$ and $\max(r(o_i(s)), 0)$ for every statement s for which both trees have no `?` opinion.

Jaccard distance The Jaccard distance has been used by [11] as the basis for calculating the similarity of process models. We apply it in the following form:

$$d(t_1, t_2) = \frac{|\text{set}(t_1) \triangle \text{set}(t_2)|}{|\text{set}(t_1) \cup \text{set}(t_2)|} \tag{6}$$

where the functions "set" and "keep" are chosen from

$$\text{set}_a((o, s)) = \text{set}_o((o, s)) \cup \text{set}_s((o, s)) \tag{7}$$
$$\text{set}_o((o, s)) = \{(x, y) \mid (x, y) \in o \wedge \text{keep}(y)\} \tag{8}$$
$$\text{set}_s((o, s)) = s \tag{9}$$
$$\text{set}_{s'}((o, s)) = \text{simple}(s) \tag{10}$$
$$\text{keep}_a(x) = 1 \tag{11}$$
$$\text{keep}_t(x) = \begin{cases} 1 & x \in \{\texttt{+}, \texttt{(+)}, \texttt{0}, \texttt{(-)}, \texttt{-}\} \\ 0 & \textit{otherwise} \end{cases} \tag{12}$$

If "set_o" is used for "set", argument weights are completely ignored; "set_s" completely ignores opinions and only looks at argument and position weights. "keep" determines if unknown (`?`) and "no opinion"s (`∅`) are included.

The argumentation software Carneades [8] uses a special case of this distance function with set = $\text{set}_{s'}$, which means that the relative number of different opinion tendencies is counted.

***p*-metric** This distance function is based on the p-metric for fuzzy sets [20].

$$d(t_1, t_2) = \left(\sum_{s \in S} ds(o_1(s), o_2(s)) + \sum_{a \in A} da(s_1(a), s_2(a)) \right)^{\frac{1}{p}} \tag{13}$$

with $p \in \mathbb{N}$, and ds, da one of

$$ds_w(o_1, o_2) = \begin{cases} 0 & \text{if } o_1 = o_2 \\ 1 & \text{if } o_1 \text{ or } o_2 \text{ in } \{\texttt{∅}, \texttt{?}\} \\ |r(o_1) - r(o_2)|^p & \text{otherwise} \end{cases} \tag{14}$$
$$ds_s(o_1, o_2) = |d_w(\text{simple}(o_1), \text{simple}(o_2))|^p \tag{15}$$
$$da_0(s_1, s_2) = 0 \tag{16}$$
$$da_s(s_1, s_2) = |w(s(s_1)) - w(s(s_2))|^p \tag{17}$$

Soergel distance This distance function is also known as weighted Jaccard distance, which has also been used by [16]. We use the following definition, which uses the same vector representation as defined for the Cosine distance above:

$$d(t_1, t_2) = 1 - \frac{\sum_i \min(V_{1_i}, V_{2_i})}{\sum_i \max(V_{1_i}, V_{2_i})} \tag{18}$$

where V_{1_i} is the i-th component of the vector representation of t_1.

VAA distance In many Voting-Advice Applications (VAAs), the distance between a user's attitudes and political party's attitudes on political positions are compared. One possibility is using proximity voting logic [17], optionally weighted, which doubles the influence of a position (as, for example, used by the German Wahl-O-Mat application [13]). We adapted the idea to our model:

$$d(t_1, t_2) = \sum_{p \in P} u(o_1(p), o_2(p)) \cdot v_{t_1,t_2}(s_1(p), s_2(p)) \cdot z(o_1, o_2) \tag{19}$$

with

$$u(o_1, o_2) = \begin{cases} 2 & \text{if } o_1 \text{ or } o_2 \text{ in } \{\texttt{+}, \texttt{-}\} \\ 1 & \text{otherwise} \end{cases} \tag{20}$$

$$v_{t_1,t_2}(s_1, s_2) = \begin{cases} 2 & \text{if } s_1 \text{ or } s_2 \text{ is in the top half (rounded down)} \\ & \quad \text{of the ratings for positions} \\ 1 & \textit{otherwise} \end{cases} \tag{21}$$

$$z(o_1, o_2) = \begin{cases} 0 & \text{if } o_1 \text{ or } o_2 \text{ in } \{\emptyset, \texttt{?}\} \\ |r(\text{simple}(o_1) - r(\text{simple}(o_2))| & \text{otherwise} \end{cases} \tag{22}$$

Note that, as in a VAA, only positions are considered, and statements which are no positions are ignored. Moreover, both arguments and positions can contain weights in our model, whereas a VAA typically only allows voters to input weights.

Weighted argumentation tree distance (WATD) In [3], we have suggested a pseudometric for argumentations with weighted edges and nodes. This metric respects the structure of an argumentation tree by limiting the influence of each branch to its importance, and giving statements deeper in the tree a lower weight. Adapted to the tree model in this paper, the metric is defined for two trees $t_1 = (o_1, s_1)$, $t_2 = (o_2, s_2)$ as follows:

$$d(t_1, t_2) = (1 - \alpha) \sum_{s \in S} \alpha^{\text{de}(s)} \left| \prod_{a \in A_{s \to I}} w(s_1(a)) r(o_1(s)) - \prod_{a \in A_{s \to I}} w(s_2(a)) r(o_2(s)) \right| \tag{23}$$

with $\alpha \in (0, 1)$ (a lower α emphasizes opinion on statements closer to the root, similar to N in the Bhavsar distance), $A_{s \to I}$ the set of all arguments from statement s to the root I, and $\mathrm{de}(s)$ the depth of a statement s, where positions have a depth of 1. This basic idea is to multiply each opinion value of t_1 with the product of all weight from the root node I to that opinion calculate the distance to the same value in t_2. Thereby, opinion difference closer to the root have a higher influence than "deeper" opinions.

4 Comparison with a Human Baseline

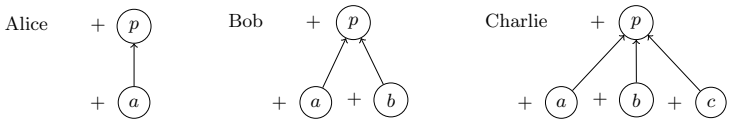
We think that the best way to check whether a distance function is intuitive is comparing it with a human baseline. In an online survey we have previously conducted [5], different possible properties for distance functions comparing attitudes in argumentation settings have been checked for their intuitiveness. In the survey, around 40 assessments by untrained human subjects have been collected for different argumentation scenarios. From the survey results, we can get a list of properties which should be fulfilled by a distance function to match human intuition. If we look only at properties which can be considered intuitive from that survey on a significance level $\alpha = 10\%$[3], we get a list of 17 properties which should be fulfilled.

For many hypotheses, also comparison questions not directly relevant for the hypotheses have been asked in the original questionnaire[4]. For instance, if we wanted to know whether Alice's attitude is more similar to Charlie's or Bob's attitude, we also asked whose attitude is closest to Bob's. For those hypotheses, we also considered properties which can be derived from the additional questions, if they are significant. Those additional properties will be marked with a superscript $^{\mathrm{A}}$, and all resulting sub-hypotheses are numbered with the according sub-question number (e.g., H2.1$^{\mathrm{A}}$ is the first question for the questionnaire scenario for H2).

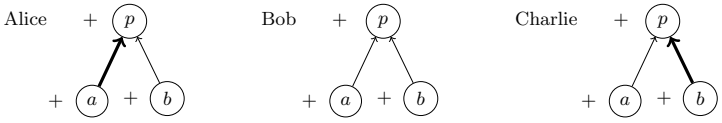
Table 2 lists all relevant hypotheses from [5] which we used as the basis for our comparison. For this paper, we changed the formulation of the hypotheses to match the real outcome of the survey to reflect the actual property expected from a distance function. Note that for H18, two scenarios were used, where only one yielded significant results, which is why this hypothesis has been completely reformulated. Figure 1 depicts visualizations of the concrete questionnaire scenarios for some more complex hypotheses, and also what similarity order is expected, based on the survey results. For example, in Figure 1e, Bob's attitude should have a smaller distance to Alice's attitude than to Charlie's attitude. Since the answers for human intuition are known only for those concrete scenarios, we will only use those concrete examples as the basis for the comparison of distance measures.

[3] including Bonferroni correction for multiple comparisons, i.e. we assure that the type I error rate is less than 10% by requiring p-values less than $\frac{\alpha}{\text{number of possible answers}}$

[4] cf. raw data at `https://github.com/hhucn/argumentation-similarity-survey-results/`

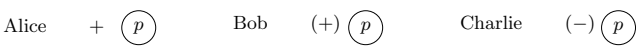


(a) H2.1$^{\mathrm{A}}$: $d(A,B) < d(A,C)$, H2.2: $d(B,C) < d(B,A)$, H2.3$^{\mathrm{A}}$: $d(C,B) < d(C,A)$

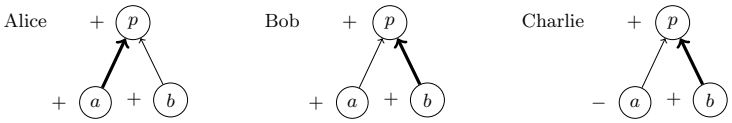


(b) H5: $d(A,B) < d(A,C)$

(c) H8.1$^{\mathrm{A}}$: $d(A,B) = d(A,C)$, H8.2$^{\mathrm{A}}$: $d(B,A) = d(B,C)$, H8.3: $d(C,A) = d(C,B)$



(d) H13.1$^{\mathrm{A}}$: $d(A,B) < d(A,C)$, H13.2: $d(B,A) < d(B,C)$, H13.3$^{\mathrm{A}}$: $d(C,B) < d(C,A)$



(e) : H14.1$^{\mathrm{A}}$: $d(A,B) < d(A,C)$, H14.2: $d(B,A) < d(B,C)$

Alice:	Bob:	Charlie:
1. $b+$	1. $a+$	1. $a+$
2. $a+$	2. $c+$	2. $b+$
3. $c+$	3. $b+$	3. $c+$

(f) H16: $d(C,B) < d(C,A)$

Alice:	Bob:	Charlie:
1. $a+$	1. $a+$	1. $c+$
2. $b+$		2. $b+$
3. $c+$		

(g) H21.1: $d(A,B) < d(A,C)$, H21.2$^{\mathrm{A}}$: $d(B,A) < d(B,C)$

Fig. 1: Visualization of questionnaire scenarios (and thus, test scenarios) of some hypotheses; $d(A,B)$ denotes the distance between Alice and Bob etc.

Table 2: All relevant hypotheses which we included in our comparison. Deviations from the original formulations in [5] are *emphasized*.

#	Property
H2	Proportionally bigger overlap on arguments for/against a position results in greater similarity than the absolute number of differences.
H3	A neutral opinion is between a positive and a negative opinion.
H5	Weights of arguments have an influence even if they are the only difference.
H7	No opinion *has the same distance from* a positive and a negative opinion *if a decision is forced*.
H8	An unknown opinion *has the same distance to a positive and a negative opinion as* a positive and a negative opinion *if a decision is forced*.
H12	It is possible for a difference in arguments for/against positions to result in greater dissimilarity than a difference in opinions on those positions.
H13	Two argumentations with weak and contrary opinions on a statement can *not* be closer than two argumentations with the same opinions, but with very different strength.
H14	Two argumentations with weak arguments and contrary opinions on their premises can *not* be closer than two argumentations with the same opinions, but with very different strength of arguments.
H16	Flipping the two most important positions results in a bigger difference than flipping two less important positions.
H18	*Moving the least important position to the top results in greater dissimilarity than changing the order of item 2 to 4.*
H19	Agreeing with someone's most important position is as important as having that person's most important opinion matching mine.
H20	Adding another most important position *(which is neutral in the other argumentations)* results in greater dissimilarity than flipping the priorities of two positions.
H21	Having more similar priorities of opinions can result in greater similarity even with lower absolute number of same opinions.
H22	Not mentioning a position results in greater dissimilarity than assigning lower priorities.

The following hypotheses have not been considered although our inclusion criterion is fulfilled: A variant of H8, which says an unknown opinion vs. a positive and a negative opinion cannot be assessed, has been excluded, because this would result in a partially defined distance function, which we consider undesirable. H9 only checked text comprehension and has no implications for a distance function. H15, which included an undercut attack, is not included since the original question was probably misleading/not understood by the participants, as discussed in [5].

We now present which distance functions fail on which reference scenarios, and give explanations on why certain distance functions fail on specific cases. We have tested each distance function with every possible combination of hyperparameters with the relevant scenarios. Table 3 summarizes which cases yield the expected results for each distance function with the best parametrization (i.e. maximum number of expected results). Those parametrisations are depicted in Table 4.

The p-metric fails on H21.1 (cf. Figure 1g) only, which happens because the missing weights for b and c have a greater influence than the common most important position of Alice and Bob. The Jaccard distance function also fails on H21.1 since it only considers that Alice and Charlie have more positions in common.

Table 3: Overview of cases fulfilled by the individual distance functions for the parametrisation which yield the highest number of fulfilled cases; ^{e} are failing cases where the calculated distance is 0; the numbers of sub-hypotheses refer to the question number in the original questionnaire.

Hypothesis	Bhavsar	Cosine	Jaccard	p-metric	Soergel	VAA	WATD
H2.1^{A}	✓	✓	✓	✓	✓	^{e}	✓
H2.2	✓	✓	✓	✓	✓	^{e}	✓
H2.3^{A}	✓	✓	✓	✓	✓	^{e}	✓
H3	✓	✓	✓	✓	✓	✓	✓
H7	✓	✓	✓	✓	✓	✓	✓
H8.1^{A}			✓	✓			
H8.2^{A}			✓	✓			
H8.3	✓	✓	✓	✓	✓	✓	✓
H5	^{e}	✓	^{e}	✓	✓	^{e}	✓
H12	✓	✓	✓	✓	✓		✓
H13.1^{A}	✓	✓	✓	✓	✓	✓	✓
H13.2	✓	✓	✓	✓	✓	✓	✓
H13.3^{A}	^{e}	✓	^{e}	✓	✓	✓	✓
H14.1^{A}	✓	✓	✓	✓	✓	^{e}	✓
H14.2	✓	✓	✓	✓	✓	^{e}	^{e}
H16	^{e}	✓	^{e}	✓	✓	^{e}	✓
H18	^{e}	✓	^{e}	✓	✓	^{e}	✓
H19	✓	✓	✓	✓	✓	✓	✓
H20	✓	✓	✓	✓	✓	✓	✓
H21.1	✓	✓			✓	^{e}	✓
H21.2^{A}	✓	✓	✓	✓	✓	^{e}	✓
H22	✓	✓	✓	✓	✓	^{e}	✓
$\sum$	16	20	17	21	20	8	19

Table 4: Best parametrisations for each distance function; each combination of the listed parameters yields the same (best) results.

Function	Best parametrisations
Bhavsar	$w \in \{w_{\bar{h}}, w_{\bar{g}}\}$, $N \in \{.1, .25, .5, .75, .9\}$
Cosine	$w \in \{w_g, w_{\bar{g}}, w_{\bar{h}}\}$
Jaccard	set $\in \{\text{set}_{s'}\}$, keep $\in \{\text{keep}_t\}$
p-metric	$p \in \{1, 2\}$, $ds \in \{d_{s_w}\}$, $da \in \{da_s\}$, $w \in \{w_{\bar{g}}, w_{\bar{h}}\}$
Soergel	$w \in \{w_g\}$
VAA	–
WATD	$\alpha \in \{.25, .5, .75, .9\}$, $w \in \{w_{\bar{g}}, w_{\bar{h}}\}$

All cases for H2 (cf. Figure 1a) fail only for the VAA distance function, where equal distances are calculated instead of different ones, because the arguments, which are the only difference in this case, are completely ignored by this function. The same applies to H5, H12, H14.1^{A} and H14.2. H14 also fails for WATD, because the distance function has been designed to *not* fulfill this property [3, Desideratum 7].

Many properties involving changing the importance order of positions, namely H16, H18, H21.1, H21.2^{A}, and H22, fail for the VAA function since it does not have a fine-grained differentiation of importance which is necessary to capture the differences.

All distance measures except for Jaccard and p-metric fail to give a positive and a negative opinion a distance which is equal to the distance to an unknown opinion (H8.1^{A}, H8.2^{A}, cf. Figure 1c). Jaccard is good here because it treats any difference of opinion as equally distant; the p-metric explicitly defines every comparison with an unknown opinion as 1. On the other hand, e.g., WATD is defined to treat an unknown opinion as falling between positive and negative, and the VAA metric ignores a position if the opinion in one graph is unknown.

H5 states that the difference between argumentations should be non-zero even if argument weights are the only difference. This fails for Bhavsar by design of the metric [2, Example 2]. The best parametrisation for Jaccard ignores weights, so it also fails here. For the same reason, H16 and H18 fail for both distance functions.

H13.3^{A} checks that a negative opinion (`-`) is closer to a weak positive opinion (`(+)`) than to a stronger positive opinion (`+`). Bhasvar and Jaccard distance functions fail to see a difference here because they treat the distances between any of the opinions `-`, `(+)`, and `+` the same.

To sum up, Cosine, p-metric, and Soergel yield the best results, matching human intuition in more than 90% percent of the tested cases.

5 Discussion

From our evaluation, one gets an idea which metrics yield intuitive results for applications which compare attitudes in argumentations. Nevertheless, we want to point out some limitations of our comparison method.

Firstly, we did not have a look at bigger argumentation hierarchies, or argumentation with re-used statements (e.g. cycles). For the former, our previous survey did not give significant results, for the latter, no reference data has been collected in the survey because cycles are hard to grasp with intuition. Hence, distance functions which model those cases (e.g. the original WATD pseudometric) have a disadvantage because this feature is not considered in the comparison.

From the survey results, it is also possible to conduct properties which should *not* be fulfilled. There are cases where there is no significant "true" answer, but there are clear "false" answers. Furthermore, the list of properties and cases checked in this paper is probably incomplete and can be extended with additional intuitive properties, which might then change the ranking of distance functions.

As we built upon the results of our previous survey, and we are not aware of similar surveys, we did not include more properties.

Note that we did not check whether the original properties as presented in Table 2 are fulfilled in general, but only whether the concrete questionnaire scenarios yielded the expected, "intuitive" results. We did this because the original survey did not find out whether the hypotheses are true, but only collected results for the specific scenarios. Moreover, all properties get equal weight. Depending on the application (e.g., a VAA), some properties might not be relevant. What is more, some distance functions might get better results if the underlying representation model is changed.

Finally, it will be interesting to evaluate distance functions not on concrete artificial scenarios, but in an application context, e.g. a recommender system, since this might produce different results. A challenge for real applications is retrieving the necessary pieces of information from a user, e.g. how important an argument is considered, within an intuitive user interface.

6 Related Work

There is only limited related research in the evaluation and development of distance function in the context of argumentation, but there are some applications of such distance functions which have been studied.

A dataset with 16 positions on 4 issues has been published by [16]. 309 students gave their opinions on those issues by giving arguments and their level of agreement with that argument on a scale from -1 (total disagreement) to 1 (total agreement). They compare different algorithms for predicting user opinions on positions. A kind of soft cosine measure, where feature similarity is exploited using position correlation, performed best in their comparison. The comparison also included, i.a., collaborative filtering using Jaccard similarity, ordinary Cosine similarity, and other, model-based algorithms, e.g. a neural network.

Their work focuses on the application of measuring similarity in the concrete context of a recommender system, whereas we focus on calculating relative similarities to get a similarity order for user attitudes. Similarly, [18] tested different recommender agents in laboratory argumentation settings. [9] uses collaborative filtering and clustering in a social network context to find political parties closest to a user. The collaborative filtering was used to predict missing values to make clustering with sparse information easier.

Related work in other domains than argumentation chose a similar way of evaluation with a human baseline as we did in this paper.

In the context of word similarity, [12] proposed different distance functions, and compared them with human ratings from a dataset created by [14]. They also indicate that the best way to determine the quality of a distance function is comparing it with human common sense. Within the same application context, [6] agrees that "comparison with human judgments is the ideal way to evaluate a measure of similarity".

The study presented in [1] is based on the study design of [14]. 50 human subjects assessed the similarity of process descriptions, and compared those assessments with the values of five metrics. The results did not correlate well, but the correlation with the metrics was not worse than the correlation between the human subjects. [11] present a metric based on the Jaccard coefficient for process model similarity. They compared the results of the metric with human assessment in an information retrieval task.

[19] evaluated six different similarity measures (i.a., $l1$, $l2$ norm, pointwise mutual information) with the application in a recommender system for online communities using item-based collaborative filtering. A similarity measure has been considered good if the user wanted to join the suggested community. The $l2$ norm performed best, although the authors found other tested measures, which incorporated mutual information, more intuitive.

7 Conclusion and Future Work

We have presented several distance functions for comparing the attitudes of different persons in an argumentation. We compared the performance of the functions in various scenarios with a human baseline taken from a survey we have previously conducted [5]. The distance functions based on the p-metric, Cosine, and Soergel distance performed best on our dataset. Those results can be used for developing applications which should give results matching human intuition, e.g. when developing a distance-based recommender system for arguments, or clustering of opinions.

For future work, an extended comparison with more scenarios for a human baseline would be useful, i.a. for deeper argumentations. A comparison in different application scenarios can give more insights. We plan to compare different metrics in an argument-based voting advice application in an empirical study. Another aspect for further research is the question of how to gather the information needed from users without having user interfaces which are too crowded.

References

1. Bernstein, A., Kaufmann, E., Bürki, C., Klein, M.: How similar is it? towards personalized similarity measures in ontologies. In: Wirtschaftsinformatik 2005, pp. 1347–1366. Springer (2005)
2. Bhavsar, V.C., Boley, H., Yang, L.: A weighted-tree similarity algorithm for multi-agent systems in e-business environments. Computational Intelligence **20**(4), 584–602 (2004)
3. Brenneis, M., Behrendt, M., Harmeling, S., Mauve, M.: How Much Do I Argue Like You? Towards a Metric on Weighted Argumentation Graphs. In: Proceedings of the Third International Workshop on Systems and Algorithms for Formal Argumentation (SAFA 2020). pp. 2–13. No. 2672 in CEUR Workshop Proceedings, Aachen (Sep 2020)
4. Brenneis, M., Mauve, M.: deliberate – Online Argumentation with Collaborative Filtering. In: Computational Models of Argument. vol. 326, p. 453–454. IOS Press (Sep 2020). https://doi.org/10.3233/FAIA200530

5. Brenneis, M., Mauve, M.: Do I Argue Like Them? A Human Baseline for Comparing Attitudes in Argumentations. In: Proceedings of the Workshop on Advances In Argumentation In Artificial Intelligence 2020. pp. 1–15. No. 2777 in CEUR Workshop Proceedings, Aachen (Nov 2020)
6. Budanitsky, A., Hirst, G.: Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In: Workshop on WordNet and other lexical resources. vol. 2, pp. 2–2 (2001)
7. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence **77**(2), 321–357 (1995)
8. Gordon, T.F.: Structured consultation with argument graphs. From Knowledge Representation to Argumentation in AI. A Festschrift in Honour of Trevor Bench-Capon on the Occasion of his 60th Birthday pp. 115–133 (2013)
9. Gottipati, S., Qiu, M., Yang, L., Zhu, F., Jiang, J.: Predicting user's political party using ideological stances. In: International Conference on Social Informatics. pp. 177–191. Springer (2013)
10. Kunz, W., Rittel, H.W.J.: Issues as elements of information systems, vol. 131. Citeseer (1970)
11. Kunze, M., Weidlich, M., Weske, M.: Behavioral similarity – a proper metric. In: International Conference on Business Process Management. pp. 166–181. Springer (2011)
12. Li, Y., Bandar, Z.A., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on knowledge and data engineering **15**(4), 871–882 (2003)
13. Marschall, S.: The online making of citizens: Wahl-O-Mat. The making of citizens in Europe: New perspectives on citizenship education pp. 137–141 (2008)
14. Miller, G.A., Charles, W.G.: Contextual correlates of semantic similarity. Language and cognitive processes **6**(1), 1–28 (1991)
15. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab (November 1999), `http://ilpubs.stanford.edu:8090/422/`, previous number = SIDL-WP-1999-0120
16. Rahman, M.M., Sirrianni, J., Liu, X.F., Adams, D.: Predicting opinions across multiple issues in large scale cyber argumentation using collaborative filtering and viewpoint correlation. The Ninth International Conference on Social Media Technologies, Communication, and Informatics pp. 45–51 (2019)
17. Romero Moreno, G., Padilla, J., Chueca, E.: Learning VAA: A new method for matching users to parties in voting advice applications. Journal of Elections, Public Opinion and Parties pp. 1–19 (2020)
18. Rosenfeld, A., Kraus, S.: Providing arguments in discussions on the basis of the prediction of human argumentative behavior. ACM Transactions on Interactive Intelligent Systems (TiiS) **6**(4), 1–33 (2016)
19. Spertus, E., Sahami, M., Buyukkokten, O.: Evaluating similarity measures: A large-scale study in the orkut social network. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. pp. 678–684 (2005)
20. Xuecheng, L.: Entropy, distance measure and similarity measure of fuzzy sets and their relations. Fuzzy sets and systems **52**(3), 305–318 (1992)