Extended Abstract – 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

1 Introduction

Comparing attitudes different people or organizations have in an argumentation is often relevant and useful, e.g. for clustering using opinions mentioned in argumentations, finding a consensus, recommender systems for argumentation platforms (such as our platform *deliberate* [3], which can be used for political education), or comparing one's own attitudes and arguments with those of political parties. Especially if used for sensitive tasks like recommending a party to vote for, it is important to have a distance measure which yields intuitive results which can be understood. In previous work [4], 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 work, we compare different distance functions regarding those properties. Our goal is to provide hints for applications which kinds of distance functions best match human intuition and where and why there are differences. In our argumentation model we consider that arguments can be of different strengths, and persons can be more or less sure about their opinions, which should be considered when calculating the distance between persons.

Our contribution is the following: First, 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. We adapted different existing distance functions to use them with attitudes in argumentations. Then 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. Finally, we discuss why the distance functions fail to fulfill some properties.

2 Comparison of Distance Functions for Argumentations

Seven distance functions are included in our comparison, of which most are based on previous works in argumentation theory or related fields, and which have been adapted by us for use with our formal definition of argumentation graphs which consider strengths of arguments and statements [2]. Many functions have different hyperparameters, for which we tested different values. We included the following functions:

- Bhavsar distance [1] (originally used for match-making of agents in e-business environments)
- Cosine distance (similar to [6], who predict opinions in argumentation)
- Jaccard distance (used in [5] as basis for calculating the similarity of process models)
- p-metric for fuzzy sets [8]
- Soergel distance (also used by [6])
- VAA distance (as used in different Voting-Advice Applications with proximity voting logic [7])
- our weighted argumentation tree distance (WATD) [2]

We think 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[1] [4], different possible properties for distance functions comparing attitudes in argumentation settings have been checked for their intuitiveness. Assessments by untrained human subjects have been collected for different argumentation scenarios. From the survey results, we got 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 $\alpha = 10\%$, we get a list of 22 properties which should be fulfilled, i.a.

1. weights of arguments have an influence even if they are the only difference,
2. *no opinion* has the same distance from a *positive* and a *negative* opinion,
3. flipping the order of two most important positions results in a bigger difference than flipping two less important positions,
4. the distance between an unknown opinion and a positive (or negative) opinion is the same as between a positive and a negative opinion.

Most properties are fulfilled by the p-metric (21 properties), Cosine, and Soergel distance (20); VAA has the worst result (8). The VAA distance does badly because it cannot deal with small weight differences and does not consider deeper arguments.

Some functions fail with some properties by design, e.g. the Bhavsar distance explicitly ignores weights if they are the only difference [1, Example 2], contradicting properties 1 and 3. Property 4 is only fulfilled by the p-metric and the Jaccard distance; the former explicitly defines every comparison with an unknown opinion as 1, the latter treats any difference of opinion as equally distant. Other functions, e.g. WATD, are defined to treat an unknown opinion as falling between positive and negative opinion, which does not match the intuition of average human subjects.

From our evaluation, one gets an idea which metrics yield intuitive results for applications which compare attitudes in argumentations, although our approach has some limitations. For instance, we had no look at bigger argumentation hierarchies, as our previous survey did not give significant results for them. Thus, further research is needed.

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

References

1. 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)
2. 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)
3. 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
4. Brenneis, M., Mauve, M.: Do I Argue Like Them? A Human Baseline for Comparing Attitudes in Argumentations. In: Proceedings of the 4th Workshop on Advances in Argumentation in Artificial Intelligence (Nov 2020), to appear
5. Kunze, M., Weidlich, M., Weske, M.: Behavioral similarity – a proper metric. In: International Conference on Business Process Management. pp. 166–181. Springer (2011)
6. 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)
7. 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)
8. Xuecheng, L.: Entropy, distance measure and similarity measure of fuzzy sets and their relations. Fuzzy sets and systems **52**(3), 305–318 (1992)