“PageRank” for Argument Relevance

Henning Wachsmuth and **Benno Stein** and **Yamen Ajjour**
Faculty of Media, Bauhaus-Universität Weimar, Germany
{henning.wachsmuth, benno.stein, yamen.ajjour}@uni-weimar.de

Abstract

Future search engines are expected to deliver pro and con arguments in response to queries on controversial topics. While argument mining is now in the focus of research, the question of how to retrieve the *relevant* arguments remains open. This paper proposes a radical model to assess relevance objectively at web scale: the relevance of an argument's conclusion is decided by what other arguments reuse it as a premise. We build an argument graph for this model that we analyze with a recursive weighting scheme, adapting key ideas of PageRank. In experiments on a large ground-truth argument graph, the resulting relevance scores correlate with human average judgments. We outline what natural language challenges must be faced at web scale in order to stepwise bring argument relevance to web search engines.

1 Introduction

What stance should I take? What are the best arguments to back up my stance? Information needs of people aim more and more at arguments in favor of or against a given controversial topic (Cabrio and Villata, 2012b). As a result, future information systems, above all search engines, are expected to deliver pros and cons in response to respective queries (Rinott et al., 2015). Recently, argument mining has become emerging in research, also being studied for the web (Al-Khatib et al., 2016a). Such mining finds and relates units of arguments (i.e., premises and conclusions) in natural language text, but it does not assess what arguments are *relevant* for a topic. Consider the following arguments (with implicit conclusions) for a query “reasons against capital punishment”:

Example a_1. *“The death penalty legitimizes an irreversible act of violence. As long as human justice remains fallible, the risk of executing the innocent can never be eliminated.”*

Example a_2. *“Capital punishment produces an unacceptable link between the law and violence.”*[1]

While both arguments are on-topic, a_1 seems more clear, concrete, and targeted, potentially making it more relevant. First approaches to assess argument quality exist (see Section 2). However, they hardly account for the problem that argument quality (and relevance in particular) is often perceived subjectively. Whether a_3, e.g., is more relevant than a_1 or less depends on personal judgment:

Example a_3. *“The death penalty doesn't deter people from committing serious violent crimes. The thing that deters is the likelihood of being caught and punished.”*

In this paper, we study from a retrieval perspective how to assess argument relevance *objectively*, i.e., without relying on explicit human judgments. Following argumentation theory, we see relevance as a dialectical quality that depends on how beneficial all participants of a discussion deem the use of an argument for the discussion (Walton, 2006). In the context of web search, an objective assessment hence at best takes place at web scale. We propose the radical model that relevance is not decided by the content of arguments, but structurally by how many arguments across the web use the conclusion of an argument as a premise and by how relevant these are in turn. The rationale is that an author cannot control who “cites” his or her argument in this way, so each citation can be assumed to add to relevance. Thereby, we achieve to decouple relevance from the soundness of the inference an argument makes to draw its conclusion.

[1] All example arguments in Sections 1–4 are derived from www.bbc.co.uk/ethics/capitalpunishment/against_1.shtml.

Given algorithms that mine arguments from the web and that decide if two argument units mean the same, an argument graph can be built where nodes represent arguments and each edge the reuse of a conclusion as a premise (Section 3). Based on the graph, we devise an adaptation of the PageRank algorithm (Page et al., 1999) to assess argument relevance. Originally, PageRank recursively processes the web link graph to infer the objective relevance of each web page from what other pages link to that page. In an according manner, we process the argument graph to compute a score for each argument unit. An argument's relevance then follows from the scores of its premises (Section 4). Analogue to the supportive nature of web links, our new *PageRank for argument relevance* counts any use of a conclusion as a premise as a support of relevance. In principle, balancing support and attack relations would also be possible, though.

At web scale, mining arguments from natural language text raises complex challenges. Since not all have been solved reliably yet, we here derive an argument graph from the complete *Argument Web* (Bex et al., 2013), a large ground-truth database consisting of about 50,000 argument units. This way, we can evaluate PageRank without the noise induced by mining errors. Moreover, we provide a first argument relevance benchmark dataset, where seven experts ranked arguments for 32 conclusions of general interest (Section 5). On the dataset, the PageRank scores beat several intuitive baselines and correlate with human average judgments of relevance—even though they ignore an argument's content and inference—indicating the impact of our approach (Section 6). We discuss how to bring argument relevance to web search engines, starting from the technologies of today (Section 7).

Contributions To summarize, the work at hand provides three main contributions to research:

1. An approach to structurally and hence objectively assess argument relevance at web scale.
2. A first benchmark ranking dataset for the evaluation of argument relevance assessment.
3. Evidence that argument relevance depends on the reuse of conclusions in other arguments.

2 Related Work

Argument relevance can be seen as one dimension of argumentation quality. In argumentation theory, two relevance types are distinguished: *Local* relevance means that an argument's premises actually help accepting or rejecting its conclusion. Such relevance is one prerequisite of a cogent argument, along with the accepability of the premises and their sufficiency for drawing the conclusion (Johnson and Blair, 2006). Here, we are interested in an argument's *global* relevance, which refers to the benefit of the argument in a discussion (Walton, 2006): An argument is more globally relevant the more it contributes to resolving an issue (van Eemeren, 2015). While Blair (2012) deems both types as vague and resisting analysis so far, we assess global relevance using objective statistics.

In (Wachsmuth et al., 2017), we comprehensively survey theories on argumentation quality as well as computational approaches to specific quality dimensions. Among the latter, Persing and Ng (2015) rely on manual annotations of essays to predict how strong an essay's argument is—a naturally subjective and non-scalable assessment. For scalability, Habernal and Gurevych (2016) learn on crowdsourced labels, which of two arguments is more convincing. Similar to us, they construct a graph to rank arguments, but since their graph is based on the labels, the subjectivity remains. This also holds for (Braunstain et al., 2016) where classical retrieval and argument-related features serve to rank argument units by the level of support they provide in community question answering.

More objectively, Boltužić and Šnajder (2015) find popular arguments in online debates. However, popularity alone is often not correlated with merit (Govier, 2010). We additionally analyze dependencies between arguments—like Cabrio and Villata (2012a) who classify attack relations between debate portal arguments. From these, they derive accepted arguments in the logical argumentation framework of Dung (1995). Relevance and acceptability are orthogonal dimensions: an argument may be relevant even if far from everyone accepts it. While probabilistic extensions of Dung's framework exist (Bistarelli et al., 2011; Dondio, 2014), they aim at the probability of logical truth. In contrast, relevance reflects the importance of arguments, for which we take on a retrieval view.

In information retrieval, relevance represents a fundamental concept, particularly in the context of search engines. A web page is seen as relevant for a search query if it contains information the querying person was looking for (Croft et al., 2009). To assess argument relevance objectively, we adapt a core retrieval technique, recursive link analysis. Due to its wide use, we build upon Google's original PageRank algorithm (Page et al., 1999), but alternatives such as (Kleinberg, 1999) would also apply. PageRank is sensitive to certain manipulations, such as link farms (Croft et al., 2009). Some of them will affect argument graphs, too. Improvements of the original algorithm should therefore be taken into account in future work. In this paper, we omit them on purpose for simplicity and clarity.

We already introduced our PageRank approach in (Al-Khatib et al., 2016a), but we only roughly sketched its general idea there. Recursive analyses have also been proposed for fact finding, assuming that trustworthy web pages contain many true facts, and that true facts will be found on many trustworthy web pages (Yin et al., 2007; Galland et al., 2010). Pasternack and Roth (2010) model a user's prior knowledge in addition. Close to argumentation, Samadi et al. (2016) evaluate claims using a credibility graph derived from evidence found in web pages. All these works target truth. In order to capture relevance, we base PageRank on the reuse of argument units instead.

In particular, we construct a graph from all arguments found in web pages. Both complex argument models from theory (Toulmin, 1958; Reisert et al., 2015) and simple proprietary models (Levy et al., 2014) have been studied for web text. Some include quality-related concepts, such as evidence types (Al-Khatib et al., 2016b). Others represent the overall structure of argumentation (Wachsmuth et al., 2015). Like Mochales and Moens (2011), we consider only premises and conclusions as units of arguments here. This is the common ground of nearly all argument-level models, and it will allow an integration with approaches to analyze argument inference (Feng and Hirst, 2011) based on the argumentation schemes of Walton et al. (2008). Edges in our graph emerge from the usage of units in different arguments. Alternatively, it would be possible to mine support and attack relations between arguments (Park and Cardie, 2014; Peldszus and Stede, 2015).

Not all mining steps work robustly on web text yet (Al-Khatib et al., 2016a). To focus on the impact of PageRank, we thus rely on ground-truth data in our experiments. In isolation, existing argument corpora do not adequately mimic web context, as they are small and dedicated to a specific genre (Stab and Gurevych, 2014), or restricted to flat relations between units (Aharoni et al., 2014). To maximize size and heterogeneity, we here refer to the Argument Web (Bex et al., 2013), which is to our knowledge the largest ground-truth argument database available so far. It includes relation-rich corpora, e.g., AraucariaDB (Reed and Rowe, 2004), as well as much annotated web text, e.g., from (Walker et al., 2012) and (Wacholder et al., 2014). Thus, it serves as a suitable basis for constructing an argument graph.

3 The Web as an Argument Graph

We now present the model that we envision as the basis for argument relevance in future web search, targeting information needs of the following kind: *"What are the most relevant arguments to support or attack my stance?"* The model relies on three principles that aim at the separation of concerns:

I. *Freedom of Inference.* No inference from argument premises to conclusions is challenged.

II. *Freedom of Mining.* No restrictions are made for how to mine and relate argument units.

III. *Freedom of Assessment.* No graph processing method is presupposed to assess relevance.[2]

3.1 Definition of the Argument Graph

Let $D = \{d_1, d_2, \ldots\}$ be the set of all considered web pages. Each $d \in D$ may contain zero or more arguments. Given D, we model the web as an argument graph in three incremental building blocks:

Canonical Argument Structure A canonical structure that represents each argument in D as a tuple $a = \langle c, P \rangle$ where c denotes the conclusion of a and $P = \{c_1, \ldots, c_k\}$ its premises, $k \geq 0$. Both conclusions and premises form argument units.

Reuse Interpretation Function An interpretation function $\mathcal{I}$ that assigns one label from the set $\{\text{“}\approx\text{”}, \text{“}\not\approx\text{”}\}$ to each pair of argument units (c, c') from all arguments in D.

Argument Graph A graph $G = (A, E)$ such that

$A = \{a_1, \ldots, a_n\}$ is a set of nodes where each $a \in A$ corresponds to one argument in D;

$E \subseteq A \times A$ is a set of edges where $(a, a') \in E$ iff. $\mathcal{I}(c, c_i) = \text{“}\approx\text{”}$ holds for the conclusion c of a and any premise c_i of a'.[3]

[2] We will introduce a specific method in Section 4. Nevertheless, other methods would also be applicable in principle.

[3] In order to keep the definition of the argument graph simple, we include *all* possible pairs of arguments for the edges here. In practice, some pairs should rather be excluded in order to counter manipulation, e.g., those *within* a web page.

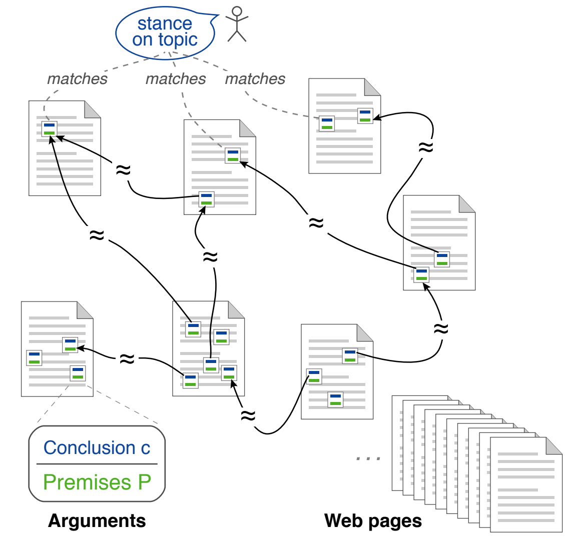

Figure 1: A small argument graph with three potentially relevant arguments for a queried stance.

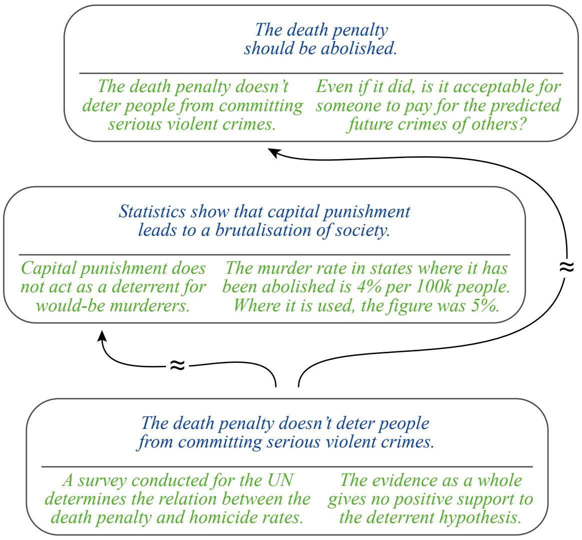
The death penalty should be abolished.
The death penalty doesn't deter people from committing serious violent crimes.
Even if it did, is it acceptable for someone to pay for the predicted future crimes of others?
Statistics show that capital punishment leads to a brutalisation of society.
Capital punishment does not act as a deterrent for would-be murderers.
The murder rate in states where it has been abolished is 4% per 100k people. Where it is used, the figure was 5%.
≈
The death penalty doesn't deter people from committing serious violent crimes.
A survey conducted for the UN determines the relation between the death penalty and homicide rates.
The evidence as a whole gives no positive support to the deterrent hypothesis.

Figure 2: Example for the reuse of an argument's conclusion as a premise in two other arguments.

Figure 1 sketches an argument graph. Given a user query with a stance on a controversial topic, as shown, each argument whose (maybe implicit) conclusion c matches the stance is potentially relevant. Stance classification is outside the scope of this paper. We assess the relevance of arguments with conclusion c. The reuse of such conclusions in other arguments is exemplified in Figure 2.

3.2 Properties of the Argument Graph Model

In accordance with Principle I, the canonical structure implicitly accepts the inference that an argument draws to arrive at its conclusion. This separates soundness from relevance, reducing the latter to an argument's units. We even permit "arguments" that have no premise. The reason is that argument units can be relevant without justification (e.g., when serving as axioms for others).

In accordance with Principle II, we do not detail the semantics of the concepts that we propose to construct arguments and their relations, leaving the exact interpretation to the mining algorithms at hand. For arguments, premises and conclusions denote the common ground, and they are generally identifiable in various web texts. For relations, the definition based on the reuse of argument units actually refines previous rather vague relation models, such as (Dung, 1995)—this is possible due to the abstraction from inference.

In accordance with Principle III, we do not predefine how to assess relevance given an argument graph (and the web pages). In addition to a conclusion, e.g., its opposite can be generated (Bilu et al., 2015) to balance support and attack somehow. In general, the usage of conclusions as premises favors a monotonous assessment (the more the better), which we implement in Section 4. Note that we allow circles in the graph. This might look unwanted as it enables circular reasoning. However, not all arguments use the same inference rule (say, modus ponens). Hence, it is reasonable that they, directly or indirectly, refer to each other.

Altogether, our model defines a *framework* for assessing argument relevance. It is instantiated by concrete mining and graph processing algorithms. An analysis of argument inference should complement this, e.g., to filter out unsound arguments. Despite its framework nature, the model suggests a recursive assessment where an argument is more relevant the more relevant arguments it relates to.

4 PageRank for Argument Relevance

Given an argument graph $G = (A, E)$, we propose to assess argument relevance structurally and thus objectively. In the following, we first develop how to adapt PageRank in order to recursively compute relevance scores for all units of the arguments in A based on E. Then, we discuss how to derive the relevance of each argument from these scores.

4.1 PageRank for Conclusion Relevance

PageRank revolutionalized web search, because it introduced *"a method for rating web pages objectively and mechanically, effectively measuring the* *human interest and attention"* (Page et al., 1999). The original method assigns a high PageRank $p(d)$ to a web page d if d is linked by many other web pages with a high PageRank. This value corresponds to the probability that a web surfer, who either follows a link on a visited web page or randomly chooses a new page, enters d. In particular, based on the link graph induced by a set of web pages D, $p(d)$ is computed recursively as:

$$p(d) \;=\; (1-\alpha)\cdot\frac{1}{|D|} \;+\; \alpha\cdot\sum_i \frac{p(d_i)}{|D_i|}$$

Here, $p(d_i)$ is the PageRank score of a web page $d_i \in D$ that links to d, and D_i is the set of all web pages that d_i links to. According to the right summand, a web page linking to d contributes to $p(d)$ more the less outgoing links it has, in order to reward the focus on specific links. The left summand specifies an equal ground relevance $\frac{1}{|D|}$ for all web pages d, summing up to 1. The factor $\alpha \in [0,1]$ weights the two summands.

Based on an argument graph $G = (A, E)$, we adapt the PageRank idea in order to analogously rate the conclusion c of each argument $a \in A$ "objectively and mechanically". Recall that an edge $(a, a') \in E$ states that c is a premise in another argument $a' \in A$. Now, we assign a high PageRank $\hat{p}(c)$ to c if c serves as a premise for many conclusions c_i with high $\hat{p}(c_i)$. For this, we adjust the equation above in two ways:

Ground Relevance Originally, PageRank works on the lowest layer of the web, the link graph. This layer has no specific entry point, which is why the ground relevance of all pages $d \in D$ is the same in $p(d)$ above. Working with arguments on web pages, however, adds a new layer on top. Therefore, we start with the original PageRank as the ground relevance, i.e., we postulate that the higher $p(d)$ is, the more relevant is a conclusion c found on d by default. In order to maintain a sum of 1 for all arguments, we normalize $p(d)$ with the average number of arguments per web page. This results in the ground relevance $\frac{p(d)\cdot|D|}{|A|}$ for each c.

Recursive Relevance The author of a web page d can specify the web pages that d links to, but the author cannot control which pages eventually link to d. This contrast is a cornerstone of the original PageRank to model relevance objectively. By analogy, the author of an argument specifies which argument units to use as premises for the argument's conclusion c, but the author cannot control which arguments use c as a premise for their conclusion c_i. This contrast is a cornerstone of our "PageRank for arguments" to model relevance objectively. In order to reward a focus on specific conclusions, we normalize the impact of the relevance $\hat{p}(c_i)$ of each conclusion c_i, for which c serves as a premise, on the relevance of c by the number of premises $|P_i|$ given for c_i. This results in the contribution $\frac{\hat{p}(c_i)}{|P_i|}$ for each c_i.

Altogether, we compute the PageRank of a conclusion c that is contained in a web page d as:

$$\hat{p}(c) \;=\; (1-\alpha)\cdot\frac{p(d)\cdot|D|}{|A|} \;+\; \alpha\cdot\sum_i \frac{\hat{p}(c_i)}{|P_i|}$$

4.2 Properties of the PageRank Approach

For space reasons, we only sketch that the adapted PageRank $\hat{p}(c)$ maintains two important properties of the original PageRank (Page et al., 1999).

First, by construction, the original scores $p(d)$ of all web pages sum up to 1. The left summand of $\hat{p}(c)$ shares this sum among all arguments. The right summand ensures that the total contribution of conclusion usages is normalized with the total number of premises. Thus, the sum of all adapted PageRank scores is also 1.

Second, as the original PageRank, $\hat{p}(c)$ reflects the idea of a citation ranking: Basic conclusions that serve as a premise for many arguments get a high score. They take the role of fundamental literature, say, *"human life is valuable"* in the context of death penalty. At the other end, each conclusion of a leaf argument in the graph is assigned only its ground relevance, since it is never reused. Without citations, relevance can still be estimated based on authorship, e.g., finding an argument on the BBC page from Footnote 1 might suffice to deem it relevant. We model this by including $p(d)$ in $\hat{p}(c)$.

4.3 From Conclusion to Argument Relevance

Given a conclusion c, all arguments $\langle c, P\rangle$ compete in terms of relevance. Since each such argument has the same conclusion, its relevance needs to be derived from its premises P. Intuitively, an argument proves only as strong at its weakest premise, so the minimum premise PageRank score could govern relevance. This fits our model, as we have "outsourced" the soundness of the inference based on the premises. However, it favors arguments with few premises. In order to find the best derivation, we compare four different premise aggregation methods in Section 6:

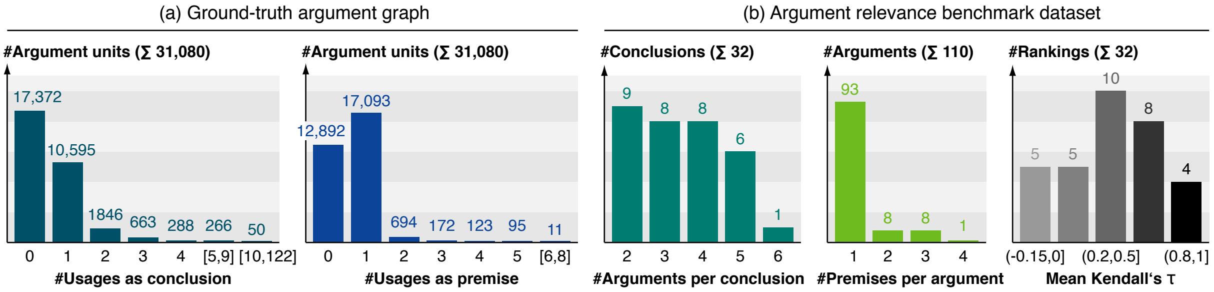

Figure 3: (a) Histograms of the usages of all argument units in the ground-truth argument graph as a conclusion or premise, respectively. (b) Histograms of the arguments per conclusion and the premises per argument in the benchmark dataset as well as the mean Kendall's τ rank correlation of all seven rankers.

(a) *Minimum.* The relevance of an argument corresponds to its minimum premise PageRank.

(b) *Average.* The relevance of an argument corresponds to its average premise PageRank.

(c) *Maximum.* The relevance of an argument corresponds to its maximum premise PageRank.

(d) *Sum.* The relevance of an argument corresponds to the sum of its premise PageRanks.

In general, as for web pages (Croft et al., 2009), PageRank should certainly not be seen as the ultimate way of assessing argument relevance, especially because it fully ignores the content and inference of arguments. Rather, it provides an objective means to identify arguments commonly referred to for a given conclusion.

5 The Webis-ArgRank-17 Dataset

This section describes our construction of a large ground-truth argument graph as well as our creation of manual relevance rankings of arguments from the graph. The resulting *Webis-ArgRank-17 dataset* is not meant for training statistical ranking approaches. Rather, it serves as a first benchmark for evaluating argument relevance assessment.[4]

5.1 A Large Ground-Truth Argument Graph

As discussed in Section 2, the *Argument Web* is the largest existing argument database. It contains structured argument corpora (several from published research) with diverse types of mostly English text, often web content.[5] The Argument Web stores annotations in a standard format, so called argument maps. Each map specifies nodes that correspond to argument units or to inference rules. Edges connect one of each in either direction. Implicitly, incoming edges of an inference node define premises of an argument, the single outgoing edge an argument's conclusion. At the last date we accessed the Argument Web (June 2, 2016), it contained 57 corpora with 8479 maps, summing up to 49,504 argument units and 26,012 arguments.

In order to get a ground-truth argument graph of maximum size, we merged all argument maps except for duplicates. We created one argument node for each inference node while maintaining argument units not connected to any inference node for completeness. For the edges of the argument graph, we assumed two units to be the same if and only if they capture exactly the same text, thereby minimizing the number of falsely detected usages of conclusions. Figure 3(a) shows how many units are used how often as a premise and as a conclusion respectively.[6]

The constructed graph contains 31,080 different argument units, 28,795 of which participate in 17,877 arguments. For convenience, we already precomputed the adapted PageRank score $\hat{p}(c)$ of each argument unit c as well as the frequency of c in the graph. As no original PageRank score $p(d)$ can be accessed for c, we started with the same ground relevance $\frac{1}{31{,}080}$ for all units.

5.2 Benchmark Argument Rankings

3113 conclusions in the constructed graph have more than one argument and, so, are candidates for ranking. From these, we selected all 498 conclusions for which at least one argument has multiply

[4] The dataset and the Java code for reproducing all experiment results are freely available at: http://www.arguana.com

[5] All corpora contained in the Argument Web can be found at: http://www.arg.dundee.ac.uk/aif-corpora

[6] We tested some high-precision heuristics to match units that occur multiple times in different manisfestations, such as ignoring capitalization or discourse connectives. However, the effect was little, which is why we decided to stick with exact matches to avoid false positives in the ground-truth graph.

Conclusion	Argument	Premises of Argument	Rank	∅
Strawberries are the best choice for your breakfast meal!	a_1 (pro)	"Berries are superfoods because they're so high in antioxidants without being high in calories", says Giovinazzo$_{\text{[premise 1]}}$ MS, RD, a nutritionist at Clay health club and spa, in New York City.$_{\text{[premise 2]}}$	1	1.43
	a_2 (pro)	One cup of strawberries, for instance, contains your full recommended daily intake of vitamin C, along with high quantities of folic acid and fiber.	2	1.57
	a_3 (pro)	Strawberries are good for your ticker.	3	3.00
Technology has enhanced the daily life of humans.	a_4 (pro)	The internet has enabled us to widen our knowledge.	1	2.00
	a_5 (pro)	Technology has given us a means of social interaction that wasn't possible before.	2	2.71
	a_6 (pro)	The use of technology has revolutionized business.	3	3.14
	a_7 (con)	No longer is shopping a personal experience, you're mostly dealing with computers when you're purchasing online.	4	3.43
	a_8 (con)	Social interactions via the internet are a huge waste of time.	5	4.29
	a_9 (con)	There's a ton of information on the internet that is entirely useless.	6	5.42

Table 1: Two argument conclusions in the benchmark dataset, together with the premises of all alternative pro and con arguments, the arguments' ranks in the dataset, and the mean ranks assigned by the 7 rankers.

used premises, as all others show no structural difference in the graph. We then let two experts from computational linguistics classify for each conclusion as to whether it denotes (a) a claim that internet users might search arguments for or (b) not such a claim for any of five reasons: (1) It is not of general interest but comes from a personal or any other too specific context, e.g., *"Viv needs to be allowed to prove herself"*, (2) its meaning is unclear, e.g., *"we need to get back to the classics"*, (3) it is not in English, (4) it mixes multiple conclusions, or (5) it is not a real conclusion but a topic, anecdote, question, or description, e.g., *"fingerprinting at the airport"* or *"what!?"*.

The experts could access the premises to see if unclear references can be resolved. They chose the same class 451 times (90.6%) with a substantial Cohen's κ agreement of 0.69. In 136 cases, no expert saw a real claim, indicating some noise in the data. For the rankings, we selected only those conclusions that both saw as claims. We disregarded multiple instances of an argument and the few conclusions where only one argument was left then.

Next, each of the 264 arguments for the remaining 70 conclusions was classified by the same experts as to whether it is (a) a correct argument for the conclusion, (b) a correct counterargument, or (c) not correct for lack of real premises. Restatements of the conclusion as well as ad-hominem attacks were not seen as premises, while the experts were asked to ignore an argument's strength.

The experts agreed in 201 cases (76.1%) with $\kappa = 0.63$. An example that they saw differently is *"I agree... the thrill is gone"* for the conclusion *"a tweet is fundamentally valueless"*. To allow for a reasonable but tractable ranking, we kept only conclusions where the experts agreed on two to six arguments and/or counterarguments, and we discarded one conclusion that paraphrased another one. The resulting dataset covers 32 conclusions. We included all 110 arguments for these conclusions, since their relevance is assessed via ranking. Figure 3(b) shows the distribution of arguments over conclusions and the premises per argument. For two conclusions, all premises are listed in Table 1. The ranks resulted from the final step.

In particular, since we expected argument relevance to be perceived subjectively, a total of seven experts from computational linguistics and information retrieval ranked all arguments (in terms of a strict ordering) for each conclusion by how much they contribute to the acceptance or rejection of the conclusion. In order not to bias the experts, they received arguments with corrected grammar, resolved references, and merged premises. They should follow their own view but acknowledge that there may be relevant counterarguments.

The highest agreement of two experts on all 32 rankings was 0.59 in terms of Kendall's τ rank correlation (Pearson coefficient 0.63), the mean over all expert pairs 0.36 (Pearson 0.40). This gap supports the subjectivity hypothesis. Figure 3(b) shows that five rankings had a negative τ-value for all experts (the lowest τ was –0.14), whereas in 12 cases τ was above 0.5, in four cases above 0.8. Overall, the resulting ranks thus largely qualify as benchmark average relevance judgments.[7]

[7] We are aware that the seven chosen experts are certainly not representative of average web users. In order to achieve a controlled setting, however, we preferred to rely on experts, e.g., to avoid misconceptions of terms such as "argument".

		(a) Minimum			**(b) Average**			**(c) Maximum**			**(d) Sum**			**(e) Best results**		
#	**Approach**	τ	*best*	*worst*	τ	*best*	*worst*	τ	*best*	*worst*	τ	*best*	*worst*	τ	*best*	*worst*
1	**PageRank**	**0.01**	**11**	**6**	0.02	**12**	**6**	0.11	**11**	**4**	**0.28**	**15**	**3**	**0.28**	**15**	3
2	Frequency	–0.10	5	9	–0.03	8	10	–0.01	7	9	0.10	11	9	0.10	11	9
3	Similarity	–0.13	7	12	–0.05	8	11	0.01	9	10	0.02	9	10	0.02	9	10
4	Sentiment	**0.01**	8	**6**	**0.11**	**12**	4	**0.12**	10	5	0.12	12	4	0.12	12	4
5	Most premises	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	0.19	6	**1**
6	Random	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	0.00	8	7

Table 2: (a–d) Mean Kendall's τ correlation of each approach with all benchmark argument rankings and counts of the *best* / *worst* rankings, once for each premise aggregation method. (e) Best observed results.

6 Evaluation

Finally, we report on an experiment that we carried out on the whole argument graph from Section 5 in order to provide first evidence that our PageRank approach objectively assesses argument relevance. Here, we assume that the average judgments of our benchmark rankings reflect objective relevance.

Approaches We compare six ranking approaches below. In case of 1.–4., we evaluate all premise aggregation methods from Section 4: (a) *Minimum*, (b) *Average*, (c) *Maximum*, and (d) *Sum*. While we are aware that more sophisticated ranking approaches are possible, the considered selection captures principle properties of arguments:

1. *PageRank.* An argument's relevance corresponds to the PageRank of its premises. This is the approach that we propose.
2. *Frequency.* An argument's relevance corresponds to the frequency of its premises in the graph. This baseline captures popularity, as proposed in related work (see Section 2).
3. *Similarity.* An argument's relevance corresponds to the similarity of its premises to its conclusion. We use the Jaccard similarity between all words in the premises and the conclusion. This basic content-oriented baseline quantifies the support of premises.
4. *Sentiment.* An argument's relevance corresponds to the positivity of its premises. Here, we sum up the positive values of all premise words in SentiWordNet (Baccianella et al., 2010) and substract all negatives. Also this baseline quantifies the support of premises.
5. *Most premises.* An argument's relevance corresponds to its number of premises. This simple baseline captures the amount of support.
6. *Random.* The relevance is decided randomly. This baseline helps interpreting the results.

Experiment For all 32 conclusions of our benchmark rankings, we assessed the relevance of every associated argument with all six approaches—in case of 1.–4. once for each premise aggregation method. For all approaches, we then compared the resulting ranks with the respective benchmark ranks and computed the mean correlation over all conclusions in terms of Kendall's τ. Kendall's τ is most suitable here, as it is meant for ranks and as it applies even when all arguments are ranked equally (unlike, e.g., the Pearson coefficient).

Results Table 2 shows that the highest rank correlation is clearly achieved by our *PageRank* approach, namely, when using the *Sum* aggregation. While a Kendall's τ of 0.28 is not very high, it can be interpreted as noncoincidental, and it is close to the low mean τ of all experts (0.36) resulting from subjectivity (see Section 5). *PageRank Sum* proves best in 15 of 32 cases. *Most premises*, which has the second highest τ (0.19), produced fewer worst rankings, but this is because it ranks all arguments equally for those 22 of the 32 conclusions where all have the same number of premises.

Matching the notion that popularity is not correlated with merit (see Section 2), *Frequency* hardly achieves anything. In fact, each frequency approach is outperformed by the PageRank approach with the respective aggregation method. The same holds for *Similarity*, which even seems to correlate rather negatively with relevance. An explanation may be that similarity rewards redundancy, which is why, e.g., all four similarity approaches falsely ranked the redundancy-free argument a_1 in Table 1 lowest. However, this requires further investigation, including an analysis of more sophisticated similarity measures. *Sentiment*, finally, performs comparably strong with $\tau > 0.1$ in three cases. In accordance with the second ranking in Table 1, this suggests that naming positive aspects (which support a conclusion) benefits relevance.

Regarding the four premise aggregation methods, we point out that their success might be partly affected by the scale of our data: 31,080 argument units is still tiny compared to the web argument graph we envision. In case of the first conclusion in Table 1, e.g., both *Mininum* and *Average* underrate the relevance of a_1, since *premise 1* fails to counter the low PageRank score of *premise 2* (resulting in $\tau = -0.82$). Summing up scores makes sense for these strongly connected premises, and it increases τ to 0.82. In contrast, *Maximum* assigns the same rank to all three arguments, which would be unlikely if web-scale data was given.

We conclude that a final judgment about our approach will require a web-scale analysis. Still, we saw first evidence for the impact of assessing argument relevance with PageRank. Considering that PageRank fully ignores an argument's content and inference—unlike our human expert rankers—its observed dominance is quite intriguing.

7 Towards Argument Search Engines

From an application viewpoint, the long-term goal of our research on argument relevance is to enable web search engines to provide the most important arguments in response to queries on controversial topics. In this regard, the proposed PageRank approach serves to retrieve relevant candidate arguments. These arguments should then be further assessed, e.g., in terms of the soundness of their inference or other quality dimensions (Wachsmuth et al., 2017). At web scale, however, our approach poses several challenges of processing natural language text, most of which refer to the construction of a reliable argument graph.

The kind of construction process that we foresee starts with the language identification and content extraction of web pages, followed by linguistic preprocessing (sentence splitting, part-of-speech tagging, etc.). For major languages, the respective technologies are not perfect but reliable (Gottron, 2008). Then, argument mining is needed in order to segment and classify argument units as well as to compose arguments. While some mining approaches for web content exist, their robustness still needs improval (Al-Khatib et al., 2016a). The most complex step is to identify the reuse of a conclusion as a premise in another argument. Ultimately, this implies that the units are semantically equivalent (or contradictory). Both textual entailment and paraphrasing help but are themselves unsolved in general. At least, promising results with about 70% accuracy are reported for ground-truth arguments (Cabrio and Villata, 2012a).

Nevertheless, the goal of bringing argument relevance to practice is not at all a dream of the far future. The decisive observation is here that the size of the web allows preferring precision over recall. In particular, an initial high-precision, lower-recall argument graph may be obtained by focusing on "low-hanging fruits". For instance, reliable arguments can be derived from those web sources that are directly cited in online debate portals, such as http://www.debatepedia.org. Generally, the mining process can be tailored to narrow domains and to well-structured text genres first. In order to limite the noise from mining errors, simple and unambiguous sentence-level arguments may be focused on and mined only if the respective approaches have a high confidence. Similarly, the recognition of equivalent argument units may be restricted to near-duplicates based on high-precision heuristics, such as ignoring capitalization, discourse connectives and other filler words, or similar.

From there on, the framework nature of the defined argument graph allows a stepwise refinement of the process, integrating new approaches to any process step as available. Research towards argument search engines can hence start now.

8 Conclusion

This paper proposes a model to integrate argument relevance in future web search, and it lays theoretical ground for research on argument relevance. In particular, we have defined how to construct an argument graph at web scale as well as how to adapt PageRank for arguments in order to objectively assess relevance given the graph. The results on our new, freely available Webis-ArgRank-17 benchmark dataset with a ground-truth argument graph of notable size suggest that PageRank outperforms both frequency-based and simple content-based relevance assessment approaches.

An evaluation at web scale is left to future work. Currently, we are working on approaches that robustly mine arguments from web pages, preferring precision over recall in order to obtain a more reliable argument graph. In general, several considerable challenges exist towards the argument search engines we envision, not only in terms of argument mining. We propose to face these challenges in order to shape the future of web search together.

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016a. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1395–1404. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016b. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Languages Resources Association (ELRA).

Floris Bex, John Lawrence, Mark Snaith, and Chris Reed. 2013. Implementing the argument web. *Communications of the ACM*, 56(10):66–73.

Yonatan Bilu, Daniel Hershcovich, and Noam Slonim. 2015. Automatic claim negation: Why, how and when. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 84–93. Association for Computational Linguistics.

Stefano Bistarelli, Daniele Pirolandi, and Francesco Santini. 2011. Solving weighted argumentation frameworks with soft constraints. In *Recent Advances in Constraints: 14th Annual ERCIM International Workshop on Constraint Solving and Constraint Logic Programming*, pages 1–18. Springer Berlin Heidelberg.

J. Anthony Blair. 2012. *Groundwork in the Theory of Argumentation*. Springer Netherlands.

Filip Boltužić and Jan Šnajder. 2015. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115. Association for Computational Linguistics.

Liora Braunstain, Oren Kurland, David Carmel, Idan Szpektor, and Anna Shtok. 2016. Supporting human answers for advice-seeking questions in CQA sites. In *Proceedings of the 38th European Conference on IR Research*, pages 129–141.

Elena Cabrio and Serena Villata. 2012a. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–212. Association for Computational Linguistics.

Elena Cabrio and Serena Villata. 2012b. Natural language arguments: A combined approach. In *Proceedings of the 20th European Conference on Artificial Intelligence*, pages 205–210.

Bruce Croft, Donald Metzler, and Trevor Strohman. 2009. *Search Engines: Information Retrieval in Practice*. Addison-Wesley, USA, 1st edition.

Pierpaolo Dondio. 2014. Toward a computational analysis of probabilistic argumentation frameworks. *Cybernetics and Systems*, 45(3):254–278.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 987–996. Association for Computational Linguistics.

Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. 2010. Corroborating information from disagreeing views. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 131–140.

Thomas Gottron. 2008. *Content Extraction — Identifying the Main Content in HTML documents*. Ph.D. thesis, Universität Mainz.

Trudy Govier. 2010. *A Practical Study of Argument*. Wadsworth, Cengage Learning, Belmont, CA, 7th edition.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599. Association for Computational Linguistics.

Ralph H. Johnson and J. Anthony Blair. 2006. *Logical Self-defense*. International Debate Education Association.

Jon M. Kleinberg. 1999. Hubs, authorities, and communities. *ACM Computing Surveys*, 31(4).

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING*

2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1489–1500. Dublin City University and Association for Computational Linguistics.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38. Association for Computational Linguistics.

Jeff Pasternack and Dan Roth. 2010. Knowing what to believe (when you already know something). In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 877–885. Coling 2010 Organizing Committee.

Andreas Peldszus and Manfred Stede. 2015. Towards detecting counter-considerations in text. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 104–109. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552. Association for Computational Linguistics.

Chris Reed and Glenn Rowe. 2004. Araucaria: Software for argument analysis, diagramming and representation. *International Journal of AI Tools*, 14:961–980.

Paul Reisert, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. 2015. A computational approach for generating Toulmin model argumentation. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 45–55. Association for Computational Linguistics.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, M. Mitesh Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence — An automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450. Association for Computational Linguistics.

Mehdi Samadi, Partha Pratim Talukdar, Manuela M. Veloso, and Manuel Blum. 2016. ClaimEval: Integrated and flexible framework for claim evaluation using credibility of sources. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 222–228.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510. Dublin City University and Association for Computational Linguistics.

Stephen E. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

Frans H. van Eemeren. 2015. *Reasonableness and Effectiveness in Argumentative Discourse: Fifty Contributions to the Development of Pragma-Dialectics*. Argumentation Library. Springer International Publishing.

Nina Wacholder, Smaranda Muresan, Debanjan Ghosh, and Mark Aakhus. 2014. Annotating multiparty discourse: Challenges for agreement metrics. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 120–128. Association for Computational Linguistics and Dublin City University.

Henning Wachsmuth, Johannes Kiesel, and Benno Stein. 2015. Sentiment flow — A general model of web review argumentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 601–611. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 812–817. European Language Resources Association (ELRA).

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

Douglas Walton. 2006. *Fundamentals of Critical Argumentation*. Cambridge University Press.

Xiaoxin Yin, Jiawei Han, and Philip S. Yu. 2007. Truth discovery with multiple conflicting information providers on the web. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1048–1052.