Aggregating Referee Scores: an Algebraic Approach

Rolf Haenni

Abstract

This paper presents a quantitative solution to the problem of aggregating referee scores for manuscripts submitted to peer-reviewed conferences or scientific journals. The proposed approach is a particular application of the Dempster-Shafer theory to a restricted setting, from which an interesting algebraic framework results. The paper investigates the algebraic properties of this framework and shows how to apply it to the score aggregation and document ranking problems. Our scheme is intended to support the paper selection process of a conference or journal, not to replace it.

1 Introduction

This paper addresses a real-world problem of quantitative judgement aggregation, one that is omnipresent in the academic world and thus of great importance to all of us. We consider the typical situation of an editor or program committee, who is in charge of evaluating the manuscripts that have been submitted to a journal or conference for publication. In a competitive setting of a scientific conference, where the maximal number of accepted papers is limited, the problem then is to select the best papers from those submitted. For this, each submission is typically sent to 3 or 4 referees, who are asked to comment and score the paper in their report. This is the core of the so-called *peer review process*, which is a well-established academic procedure to guarantee high scientific standards and to prevent the dissemination of unwarranted claims or unacceptable interpretations.

Many journals and conferences ask the referees to quantitatively score the papers by a pair of values from respective scales, one that reflects the overall[1] quality of the paper and one that indicates the referee's own level of expertise or confidence. For the selection of the best papers, the editor or program committee faces then the problem of combining those scores to establish an overall ranking, from which the highest-ranked papers are accepted. The combination of such referee scores is a simple real-world judgement aggregation problem. Most of the existing convenient on-line conference management systems (e.g. ConfMaster, ConfTool, EasyChair, Linklings, OpenConf, Paperdyne, Start V2, WebChairing, etc.) are very rich in all kind of features for effortlessly accomplishing many complicated tasks of the peer review process, but they are usually very poor in providing automated decision support tools for the aggregation and ranking of referee scores. As a consequence, the score aggregation and paper ranking problems are still being solved manually today, and due to the many aspects and parameters to be taken into account, the resulting time-consuming procedure risks at producing bad quality results in form of unfair decisions. But what would be a reasonable procedure of combining the referee scores and establishing a ranking of peer-reviewed papers automatically?

1.1 Related Work

By describing a set of so-called *process patterns*, Nierstrasz gives an informal answer to the above question [26]. Examples of such patterns are "*Group papers according to their*

[1] Some journals and conferences ask to score various quality criteria independently of each other. The method presented in this paper is compatible with such multi-criteria scores, but we not treat them explicitly.

highest and lowest score" or "*Take care to identify papers with both extreme high and low scores*". For the scores, Nierstrasz proposes four quality categories A="*Good paper*" to D="*Serious problems*" and three levels of expertise X="*I am an expert*" to Z="*I am not an expert*".[2] Notice that Nierstrasz' pattern language has become something like the de facto standard for conferences in many computer science areas, and it is implemented in the conference management systems CyberChair [33] and Continue [23], and rudimentally in Confious [27], HotCRP [22], and MyReview [28]. The success of Nierstrasz' patterns is perfectly comprehensible from the pragmatic point of view of experienced program committee members, but from the more formal perspective of an expert in quantitative judgement aggregation or reasoning and decision making under uncertainty, they give the impression of being constituted on an ad hoc basis and may therefore seem a bit rudimentary.

A partial answer to the above question can be found in the early literature on probability from the late 17th and early 18th centuries [20, 30]. At that time, studying probability was often motivated by judicial applications, such as the reliability of witnesses in the courtroom, or more generally by the credibility of testimonies on past events or miracles. The first two combination rules for testimonies were published in an anonymous article [1]. One of them considers two independent witnesses with respective credibilities (frequencies of saying the truth) p_1 and p_2. If we suppose that they deliver the same report, they are either both telling the truth with probability $p_1 \cdot p_2$, or they are both lying with probability $(1 - p_1) \cdot (1 - p_2)$. Every other configuration is obviously impossible. The ratio of truth saying cases to the total number of cases,

$$\frac{p_1 \cdot p_2}{p_1 \cdot p_2 + (1 - p_1) \cdot (1 - p_2)}, \tag{1}$$

represents then the combined credibility of both witnesses. The more general formula for n independent witnesses of equal credibility p,

$$\frac{p^n}{p^n + (1 - p)^n}, \tag{2}$$

has been mentioned by Laplace [24]. This formula is closely related to the *Condorcet Jury Theorem* discussed in social choice theory [2, 25]. Boole mentioned in [3] a similar formula that includes a prior probability of the hypothesis in question.

A recent article picks up these ancient ideas and turns them into a very general and flexible model of combining reports from partially reliable sources [15]. The generality of the model allows it to be applied to situations of incompetent or even dishonest witnesses, who may deliver highly contradictory testimonies. At its core, the model presupposes a non-additive measure of belief [13] in form of Dempster-Shafer belief functions [9, 29], but Laplace's and Boole's formulae are included as additive special cases. The model also includes various Bayesian approaches, which require a prior probability of the hypothesis in question to turn it into a corresponding posterior probability. The method discussed in this paper is another very particular case of the general model from [15].

1.2 Problem Formulation

The formal problem setting in this paper is the following. Let $\mathcal{D}$ be a set of submitted manuscripts (documents) and $\mathcal{R}$ a set of referees. We assume that the referees are anonymous and independent. The set of assigned referees for document $D \in \mathcal{D}$ is denoted by

[2] It should be emphasized here that Nierstrasz' categorization includes an explicit operational meaning, e.g. "*I will champion the paper*" for the category A="*Good paper*" and ditto for the other categories. The whole point of the pattern language is about how people will *behave* during a program committee meeting, which is why it is not directly applicable to a non-interactive setting such as journal paper reviewing. During PC meetings, however, the notion of championing can help to keep discussions focussed.

referees$(D) \subseteq \mathcal{R}$, that is *referees* $: \mathcal{D} \to \mathcal{P}(\mathcal{R})$ is a mapping from $\mathcal{D}$ to the power set of $\mathcal{R}$. Similarly, *documents*$(R) =$ *referees*$^{-1}(R) \subseteq \mathcal{D}$ denotes the set of documents assigned to referee $R \in \mathcal{R}$. If *referees*$(D) = \{R_1, \ldots, R_k\}$ is the set of referees assigned to a particular document D, then we assume to obtain a set of respective scores, *scores*$(D) = \{s_1, \ldots, s_k\}$, each of which being a pair $s_i = (q_i, e_i) \in [0,1] \times [0,1]$ of values between 0 and 1.[3] The value $q_i \in [0,1]$ is interpreted as the referee's estimate of the paper's overall quality and $e_i \in [0,1]$ as the referee's estimate of its own level of expertise, with the usual convention that higher values represent higher quality and expertise levels.

Given the above formal setting, this paper addresses the following interlinked problems of aggregating and ranking the given referee scores.

Aggregation. For each $D \in \mathcal{D}$, derive from *scores*(D) the document's combined overall score $s_D = (q_D, e_D) \in [0,1] \times [0,1]$.

Ranking. For a given set of combined overall scores, $\mathcal{S} = \{s_D : D \in \mathcal{D}\}$, determine a total preorder $\succeq$ according to which the documents in $\mathcal{D}$ are ranked (that is D_1 is preferred to D_2 iff $s_{D_1} \succeq s_{D_2}$).

The proposed process is thus a two-step procedure, in which the papers' overall scores appear as an intermediate result before the final ranking is established. Note that the ranking problem includes the decision problem of accepting/rejecting papers as a borderline case. For a single document, i.e. for $|\mathcal{D}| = 1$, we obtain another borderline case, one that corresponds to the situation of a single paper submitted to a journal.

As an alternative, it could also be assumed that each referee's set of assigned papers, *documents*$(R) \subseteq \mathcal{D}$, is first turned into a local ranking (or decision), from which the global ranking (or decision) is then established in a second step. Such a procedure is suggested in [11]. Its main advantage is the fact that scoring, with all its inherent problems such as the referees' diverging standards with judging the merits and weaknesses of each paper on a common scale, is no longer required. The whole problem of evaluating papers according to their overall quality is thus reduced to a ranking problem. Note that aggregating local rankings into a global ranking may become very difficult or even impossible if the average number of referees per paper is low. This is a consequence of the fact that rankings are usually less informative than corresponding sets of scores.

Another important advantage of the proposed two-step procedure is the possibility to repeat Step 1 with an updated set of scores. This may be necessary for papers with an unsatisfactory overall expertise level e_D. The ability to make such a distinction between papers reviewed by a group of experts from a paper reviewed by a group of non-experts is one of the main reason for the proposed 2-dimensional scores.

1.3 Overview

This paper proposes a solution to the two problems stated above. First we suggest a formal method for combining referee scores, and then we discuss a solution for projecting combined referee scores into a total order, from which the final document ranking results. The core of the suggested method is a particular application of what is known in the literature of uncertain reasoning as *Dempster's rule of combination*, a key concepts in the *Dempster-Shafer theory* (DST) of belief functions [9, 29]. Here we adopt Dempster's original interpretation of his theory as a generalization of classical probabilistic inference [10]. For this, we look at a single score $s_i = (q_i, e_i)$ of a peer-reviewed paper as a pair of respective probabilities.

[3] In practice, typical scales for referee scores are discrete sets such as $\{1, 2, \ldots, 10\}$ or $\{$*very_poor*, *poor*, *medium*, *good*, *very_good*$\}$. To make such cases compatible with our formal setting, we assume a mapping σ from the respective set into the unit interval $[0,1]$, e.g. $\sigma($*very_poor*$) = 0.1$, $\sigma($*poor*$) = 0.3$, etc.

Formally, this allows us to consider each score as a particular representation of the referee's *opinion* about the paper. Opinions are the key elements in Jøsang's theory of subjective logic [18, 19], a particular interpretation of DST. To deal with such opinions mathematically, we follow the algebraic setting proposed in [7, 17], in which the set of all possible opinions is considered as a commutative monoid with respect to Dempster's rule of combination.

In Section 2, we first give a short introduction to the Dempster-Shafer theory, and then we present the algebraic structure of the above-mentioned opinion calculus. This is the mathematical and computational foundation of the proposed score aggregation method presented in Section 3, where scores are interpreted as respective probabilities in a restricted Dempster-Shafer model. The conclusions in Section 4 close the paper.

2 The Opinion Calculus

In its original form [9, 10], the Dempster-Shafer theory proposes a particular application of probabilistic reasoning. Its main components are two sample spaces Ω and Θ, which are interlinked by a multi-valued mapping $\Gamma : \Omega \to \mathcal{P}(\Theta)$. A set $\Gamma(\omega) \subseteq \Theta$ is thus assigned to every element $\omega \in \Omega$. For a given probability space $(\Omega, \mathcal{F}, P)$, Dempster's theory shows how to carry the probability measure $P : \mathcal{F} \to [0,1]$ defined over a σ-algebra $\mathcal{F}$ of subsets of Ω into a system of *lower* and *upper probabilities* over subsets $A \subseteq \Theta$,

$$\begin{aligned} P_*(A) &= P(\{\omega \in \Omega : \Gamma(\omega) \subseteq A\} \,|\, \{\omega \in \Omega : \Gamma(\omega) \neq \emptyset\}) \\ &= \frac{P(\{\omega \in \Omega : \emptyset \neq \Gamma(\omega) \subseteq A\})}{1 - P(\{\omega \in \Omega : \Gamma(\omega) = \emptyset\})} \end{aligned} \tag{3}$$

and

$$P^*(A) = 1 - P_*(A^c) = \frac{P(\{\omega \in \Omega : \Gamma(\omega) \cap A \neq \emptyset\})}{1 - P(\{\omega \in \Omega : \Gamma(\omega) = \emptyset\})}, \tag{4}$$

for which $P_*(A) \leq P^*(A)$ holds for all $A \subseteq \Theta$. A quadruple $(\Omega, P, \Gamma, \Theta)$ is sometimes called *hint* [21] or *Dempster space* [16].

Later, Shafer introduced a non-probabilistic interpretation of the Dempster's theory and suggested *belief* and *plausibility*, denoted by $Bel(A)$ and $Pl(A)$, as replacements for lower and upper probability [29]. Shafer's viewpoint and terminology has been adopted by many other authors, e.g. by Smets in his *Transferable Belief Model* [32]. They all depart from Dempster's original model by considering belief and plausibility functions over the so-called *frame of discernment* Θ that are not necessarily induced by an underlying probability space $(\Omega, \mathcal{F}, P)$ and its link over the multi-valued mapping Γ. There is an axiomatic system for such belief and plausibility functions [31], similar to Kolmogorov's system of axioms for probabilities. If Θ is finite, belief and plausibility functions are often expressed in terms of their underlying *mass function*,

$$m(A) = P(\{\omega \in \Omega : \Gamma(\omega) = A\}), \tag{5}$$

which is additive with respect to $\mathcal{P}(\Theta)$ and thus sums up to one over all $A \subseteq \Theta$. It is obvious that $m : \mathcal{P}(\Theta) \to [0,1]$ is a true generalization of a classical probability mass function $p : \Theta \to [0,1]$, and that Bel, Pl, and m are connected as follows:

$$Bel(A) = \frac{1}{1-m(\emptyset)} \sum_{\emptyset \neq B \subseteq A} m(B), \quad Pl(A) = \frac{1}{1-m(\emptyset)} \sum_{A \cap B \neq \emptyset} m(B). \tag{6}$$

Many variations of this general scheme have been proposed in the literature, but here we prefer to strictly follow Dempster's and Shafer's original views.

2.1 Opinions

To solve the particular problem of this paper, we restrict Dempster's original probabilistic model to a very particular case. First of all, we only consider finite sample spaces Ω, which allows us to replace the σ-algebra $\mathcal{F}$ by the power set $\mathcal{P}(\Omega)$ of Ω. Second, we only consider frames of discernment Θ of size two, e.g. $\Theta = \{H, \neg H\}$, where H and $\neg H$ denote complementary outcomes. This implies that $\{\emptyset, \{H\}, \{\neg H\}, \Theta\}$ is the codomain of the multi-valued mapping Γ. The essence of the whole structure $(\Omega, P, \Gamma, \Theta)$ can then be reduced to a simple pair $(b, d) \in [0,1] \times [0,1]$ with $b = Bel(\{H\})$, $d = Bel(\{\neg H\}) = 1 - Pl(\{H\})$, and therefore $b + d \leq 1$. In [7, 16, 17], such pairs are called *Dempster pairs* and the set of all such pairs is called *Dempster domain*. Corresponding additive triplets $\varphi_H = (b, d, i)$ with $i = 1 - b - d$ are sometimes called *opinions* about H [14, 18, 19].[4] As shown in Figure 1, opinons can be depicted as respective points in an equilateral triangle (2-simplex) called *opinion triangle* [18].[5] The three coordinates represent respective degrees of *belief* (probability assigned to "*H is true*"), degrees of *disbelief* (probability assigned to "*H is false*"), and degrees of *ignorance*[6] (probability assigned to "*I don't know*"). The correct mathematical term for the geometry of this picture is *barycentric coordinates* [4].

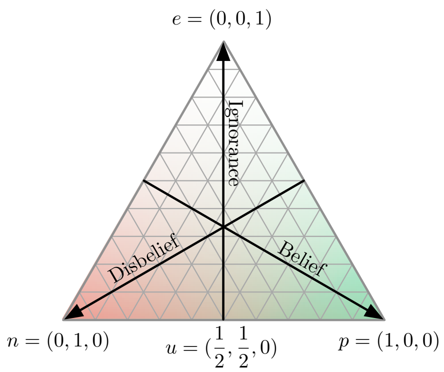

Figure 1: The opinion triangle with its three dimensions.

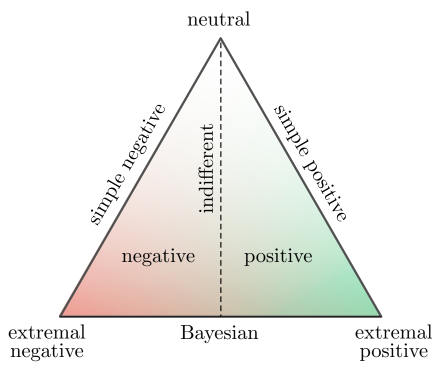

Figure 2: Various special types of opinions.

The three corners of the opinion triangle represent particular extreme cases. We adopt the terminology of [7, 17], i.e. $p = (1,0,0)$ and $n = (0,1,0)$ are called *extremal* and $e = (0,0,1)$ is called *neutral*. A general opinion (b, d, i) is called *positive* if $b > d$, and it is called *negative* if $b < d$. Positive opinions are located on the left hand side and negative opinions on the right hand side of the opinion triangle. (b, d, i) and (d, b, i) are regarded as *opposite* opinions, i.e. the opposite of a positive opinion is negative, and vice versa.

In the opinion triangle, the two regions of positive and negative opinions are separated by the central vertical line of *indifferent* opinions with $b = d$. Note that the neutral opinion e is indifferent. Other particular indifferent opinions are the points $u = (\frac{1}{2}, \frac{1}{2}, 0)$ at the bottom and $c = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ in the center of the triangle. Indifferent opinions are their own opposite.

Opinions are called *simple* (or *pure*) if either $b = 0$ or $d = 0$, i.e. $(b, 0, 1-b)$ is simple positive for $b > 0$ and $(0, d, 1-d)$ is simple negative for $d > 0$. Simple opinions are located on the left and the right edge of the triangle. Note that e, n, and p are simple. The opinions

[4] Later in [19], opinions are defined as quadruples (b, d, i, a) with an additional component a, the so-called *relative atomicity* (we do not need this in this paper).

[5] Sometimes, isosceles instead of equilateral triangles are used to visualize the Dempster domain [7, 16].

[6] In [19], Jøsang calls i degree of *uncertainty* rather than degree of ignorance, but the latter seems to be more appropriate and in better accordance with the literature.

$(b, 1-b, 0)$ at the bottom line of the triangle are called *Bayesian* (or *probabilistic*). Note that the extremal simple opinions p and n are also Bayesian, as well as the particular indifferent opinion u. All those particular types of opinions are shown in Figure 2.

2.2 Combining Opinions

One of the key components of the Demster-Shafer theory is a rule to combine two Dempster spaces $(\Omega_1, P_1, \Gamma_1, \Theta)$ and $(\Omega_2, P_2, \Gamma_2, \Theta)$ for a common frame of discernment Θ. Such a combination is usually denoted by symbols like $\otimes$ or $\oplus$ (here we prefer to use $\otimes$). If we assume P_1 and P_2 as being stochastically independent, then we naturally obtain

$$(\Omega_1, P_1, \Gamma_1, \Theta) \otimes (\Omega_2, P_2, \Gamma_2, \Theta) = (\Omega_1 \times \Omega_2, P_1 \cdot P_2, \Gamma_1 \cap \Gamma_2, \Theta) \tag{7}$$

for the combined structure [9, 21]. Translated into Shafer's terminology for two mass functions m_1 and m_2, we get what is known today as (unnormalized) *Dempster's rule of combination* or simply *Dempster's rule*:

$$m_1 \otimes m_2(A) = \sum_{B_1 \cap B_2 = A} m_1(B_1) \cdot m_2(B_2). \tag{8}$$

It is easy to see that Dempster's rule is commutative and associative, which means that the order in which opinions are combined is irrelevant. In the particular case of $\Theta = \{H, \neg H\}$ with two opinions $\varphi_1 = (b_1, d_1, i_1)$ and $\varphi_2 = (b_2, d_2, i_2)$, we know from [7, 12, 17] that Dempster's rule can be rewritten more compactly as

$$\varphi_1 \otimes \varphi_2 = \left(\frac{b_1 b_2 + b_1 i_2 + i_1 b_2}{1 - b_1 d_2 - d_1 b_2}, \frac{d_1 d_2 + d_1 i_2 + i_1 d_2}{1 - b_1 d_2 - d_1 b_2}, \frac{i_1 i_2}{1 - b_1 d_2 - d_1 b_2} \right), \tag{9}$$

and it includes (1) as special cases for $i_1, i_2 = 0$. This equation is the mathematical and computational basis for the proposed solution of the score aggregation problem in Section 3. Note that the combination of opposite extremal opinions, $p \otimes n$ or $n \otimes p$, is undefined.

2.3 The Opinion Monoid

The space of all possible opinions, $\Phi = \{(b, d, i) \in [0, 1]^3 : b + d + i = 1\}$, together with the particular form of Dempster's rule given in (9) forms an interesting algebraic structure. A thorough analysis of this structure is presented in [7, 17], where the set of all non-extremal opinions together with Dempster's rule is called *Dempster semigroup*. The extremal opinions are excluded to avoid the above-mentioned undefined combination. Here we pick up these ideas, but instead of excluding extremal opinions, we include $z = (1, 1, -1)$ as an additional opinion and call it *inconsistent*. The set of all such opinions, including the inconsistent one, is denoted by $\Phi_z = \Phi \cup \{z\}$, and $\otimes$ is extended by $p \otimes n = n \otimes p = z$. Note that z is absorbing with respect to $\otimes$, i.e. $z \otimes \varphi = \varphi \otimes z = z$ for all $\varphi \in \Phi_z$.

With this extension, the structure $(\Phi_z, \otimes)$ is closed under the operator $\otimes : \Phi_z \times \Phi_z \to \Phi_z$. From the commutativity and associativity of $\otimes$, it follows that $(\Phi_z, \otimes)$ is a *commutative semigroup*. Note that for $e = (0, 0, 1)$ we get $e \otimes \varphi = \varphi \otimes e = \varphi$ for all $\varphi \in \Phi_z$, i.e. $e = (0, 0, 1)$ is the (neutral) *identity element* of the combination. The structure $(\Phi_z, \otimes, e)$ is thus a *commutative monoid* with an absorbing *zero element* z. Since $\otimes$ is generally not invertible, $(\Phi_z, \otimes, e)$ is not a group. As is it a common practice in abstract algebra, we will refer to $(\Phi_z, \otimes, e)$ simply as Φ_z and call it the *opinion monoid*. Note that Φ_z has exactly three idempotent elements e, u, and z, i.e. $\varphi \otimes \varphi = \varphi$ only holds for $\varphi \in \{e, u, z\}$.

As pointed out in the classification of the previous subsection, Φ_z contains a number of interesting subsets. Some of them are again closed under combination andpreserve the above-mentioned algebraic properties. Table 1 gives an overview of those sub-monoids with their respective identity and zero elements. Note that a non-extremal Bayesian opinion is invertible by combining it with its own opposite, i.e. $(b, 1-b, 0)\otimes(1-b, b, 0) = u$ holds for all $b \notin \{0, 1\}$. Mathematically speaking, the set of consistent non-extremal Bayesian opinions, $\Phi_0 \setminus \{p, n, z\}$, forms an commutative (abelian) group. Note further that the set $\Phi_0 \setminus \{z\}$ possesses a natural total order $\succeq_0$ defined by $(b_1, 1-b_1, 0) \succeq_0 (b_2, 1-b_2, 0)$ iff $b_1 \geq b_2$. This order is important in Section 3 to establish the document ranking.

Name	Notation	Definition	Identity	Zero
general (extended with z)	Φ_z	$\Phi \cup \{z\}$	e	z
non-negative	$\Phi_{\geq}$	$\{(b, d, i) \in \Phi : b \geq d\}$	e	p
simple non-negative	Φ_{+}	$\{(b, d, i) \in \Phi : d = 0\}$	e	p
non-positive	$\Phi_{\leq}$	$\{(b, d, i) \in \Phi : b \leq d\}$	e	n
simple non-positive	Φ_{-}	$\{(b, d, i) \in \Phi : b = 0\}$	e	n
indifferent	$\Phi_{=}$	$\{(b, d, i) \in \Phi : b = d\}$	e	u
Bayesian (extended with z)	Φ_0	$\{(b, d, i) \in \Phi : i = 0\} \cup \{z\}$	u	z

Table 1: Different algebraic sub-structures of the opinion monoid Φ_z with their respective identity and zero elements.

2.4 Transformations

As pointed out in [7, 17], the combination of a general opinion $\varphi \in \Phi_z$ with the uniform Bayesian opinion u defines a *homomorphism* $h : \Phi_z \to \Phi_0$, i.e. $h(\varphi_1 \otimes \varphi_2) = h(\varphi_1) \otimes h(\varphi_2)$ holds for all $\varphi_1, \varphi_2 \in \Phi_z$. We can thus use h to transform a general opinion $\varphi \in \Phi_z$ into a Bayesian opinion $h(\varphi) = \varphi \otimes u \in \Phi_0$. For $\varphi = (b, d, i) \in \Phi$ we can simplify (9) into

$$h(\varphi) = \left(\frac{1-d}{(1-b)+(1-d)}, \frac{1-b}{(1-b)+(1-d)}, 0\right) = \left(\frac{1-d}{1+i}, \frac{1-b}{1+i}, 0\right), \tag{10}$$

whereas $h(z) = z$ holds as usual. A more general version of this mapping is called *plausibility transformation* and $h(\varphi)$ is called *relative plausibility* [5, 6, 8]. Note that indifferent opinions always map into u, i.e. $h(\varphi) = u$ holds for all $\varphi \in \Phi_{=}$. This includes $h(e) = u$ as a special case. In the opinion triangle, applying h to a general opinion $\varphi \in \Phi$ means to intersect the straight line through φ and z with the bottom line of the opinion triangle [7, 17]. In other words, the intersection of the opinion triangle with the straight line through $\varphi_0 \in \Phi_0 \setminus \{z\}$ and z corresponds to the preimage $h^{-1}(\varphi_0) = \{\varphi \in \Phi : h(\varphi) = \varphi_0\}$ of φ_0. This geometric interpretation of h is illustrated in Figure 3.

A similar, but non-homomorphic transformation $g : \Phi_z \to \Phi_0$ results from applying a scheme similar to the one in (10). Instead of replacing the components b and d of a general opinion $\varphi = (b, d, i)$ by respective normalized plausibilities, the idea of

$$g(\varphi) = \left(\frac{b}{b+d}, \frac{d}{b+d}, 0\right) \tag{11}$$

is to normalize b and d directly.[7] A general form of this mapping is called *belief transformation* and $g(\varphi)$ is called *relative belief* [6, 8]. Note that $g(e)$ is undefined in (11), but we can set $g(e) = u$ by default. As shown in Figure 4, applying g to $\varphi \in \Phi_z \setminus \{e\}$ means to intersect the straight line through the points φ and e with the bottom line of the opinion triangle. This geometric interpretation also holds for the special case $g(z) = u$.

[7]This transformation is a homomorphism with respect to the *disjunctive rule of combination* [7].

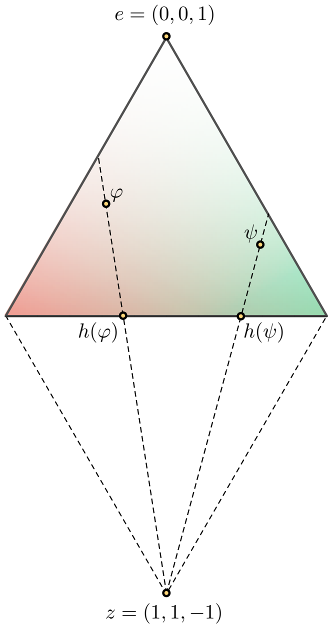

Figure 3: The homomorphic plausibility transformation h.

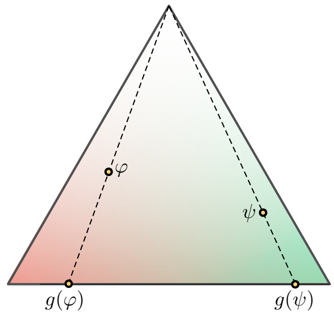

Figure 4: The belief transformation g.

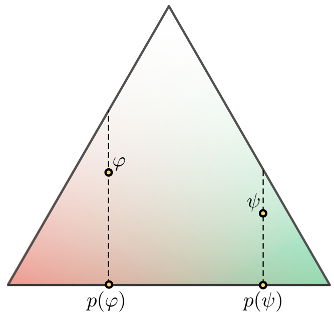

Figure 5: The pignistic transformation p.

Another interesting non-homomorphic transformation $p : \Phi_z \to \Phi_0$ redistributes the component i from a general opinion $\varphi = (b, d, i) \in \Phi_z$ equally among b and d,

$$p(\varphi) = \left(b + \frac{i}{2}, d + \frac{i}{2}, 0\right), \tag{12}$$

which includes $p(z) = u$ as a special case. In the opinion triangle, applying p to $\varphi \in \Phi$ means to project φ vertically onto the bottom line of Bayesian opinions. This is illustrated in Figure 5.[8] Note that p is a special case of what Smets calls *pignistic transformation* [32].

We prefer not to give any recommendation with regard to the question of which transformation to use, all of them have their respective advantages and disadvantages [5, 32]. In general, one can say that g tends to enforce existing degrees of belief and disbelief less cautiously than h, and that p lies somewhere in between.

3 Combining and Ranking Referee Scores

In this section, we use the algebraic investigation of the previous section as the mathematical and computational foundation for solving the problems of combining and ranking referee scores. First we show how to interpret the scores of a given document as respective opinions, one for each referee to which the document has been assigned. The independence assumption allows us then to apply the combination operator $\otimes$ defined in (9) to obtain the combined *group opinion* of all referees, from which the documents overall score is derived. Finally, we

[8] It is interesting to observe in Figure 3–5 that g, h, and p are special cases of a whole class of *symmetric transformations* obtained by intersecting the bottom line of the opinion triangle with a straight line through φ and $\varphi^i = (\frac{1-i}{2}, \frac{1-i}{2}, i)$ with $i \in \mathbb{R} \setminus [0, 1) \cup \{-\infty, +\infty\}$. In particular, we have $i = 1$ for g, $i = -1$ for h, and $i = \pm\infty$ for p.

explain how to use the proposed transformations g, h, or p to establish the final ranking, from which the highest-ranked documents are accepted.

3.1 Combining Referee Scores

As stated in Subsection 1.2, we first need to consider the problem of deriving from the set of $scores(D)$ the document's combined overall score s_D. The general idea for this is to apply Dempster's rule to corresponding opinions. Recall that a score is a point $s = (q, e) \in [0,1]\times[0,1]$ in the unit square. The obvious question now is how to map such scores into opinions. Mathematically speaking, we are looking for a meaningful mapping $\Delta : [0,1]\times[0,1] \to \Phi$, which assigns a unique opinion $\Delta(s) \in \Phi$ to each score $s \in [0,1]\times[0,1]$. We can thus look at Δ as a transformation of the unit square into the opinion triangle.

To define such a transformation, we consider each referee as a partially reliable information source. Inspired by the general model of partially reliable information source in [15], we assume that reports of unreliable sources are entirely neglected. Intuitively, this is the case whenever a referee is not an expert for reviewing a particular paper. Note that $s = (q, e)$ delivers an estimate $e \in [0,1]$ of the referee's expertise level, i.e. if we assume the referee as being trustworthy with respect to giving such an estimate, then we may interpret e as the probability $P(\{E\}) = e$ of the referee being an expert for the document's topic. Similarly, we may interpret the quality estimate q as the conditional probability $P(\{Q\}|\{E\}) = q$ of the paper being a high-quality paper, given that the referee is an expert. With $\Omega_E = \{E, \neg E\}$ and $\Omega_Q = \{Q, \neg Q\}$ we denote respective sets of outcomes. This implies a probability space $(\Omega, \mathcal{P}(\Omega), P)$ with $\Omega = \Omega_E \times \Omega_Q = \{(E,Q), (E,\neg Q), (\neg E, Q), (\neg E, \neg Q)\}$ and

$$P(\{E,Q\}) = e \cdot q, \qquad P(\{\neg E,Q\}) = (1-e) \cdot q,$$
$$P(\{E,\neg Q\}) = e \cdot (1-q), \qquad P(\{\neg E,\neg Q\}) = (1-e) \cdot (1-q).$$

Consider now another space $\Theta = \{H, \neg H\}$ and let $\Gamma : \Omega \to \mathcal{P}(\Theta)$ be defined by $\Gamma(E,Q) = \{H\}$, $\Gamma(E,\neg Q) = \{\neg H\}$, and $\Gamma(\neg E, Q) = \Gamma(\neg E, \neg Q) = \Theta$. The idea is to adopt the referee's judgment with respect Q and $\neg Q$ whenever the referee is an expert, and to discard it otherwise. This defines a Dempster space $(\Omega, P, \Gamma, \Theta)$, from which we derive $Bel(\{H\}) = e \cdot q$ and $Bel(\{\neg H\}) = e \cdot (1-q)$. Finally, this leads to the requested transformation,

$$\Delta(s) = (e \cdot q, e \cdot (1-q), 1-e), \tag{13}$$

which maps points from the unit square into the opinion triangle, as illustrated in Figure 6. Note that all scores $(q, 0)$ are mapped into the neutral opinion $(0,0,1)$, whereas all scores $(q, 1)$ are mapped into Bayesian opinions $(q, 1-q, 0)$.

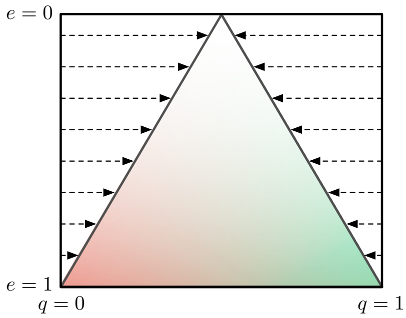

Figure 6: Transforming referee scores into opinions.

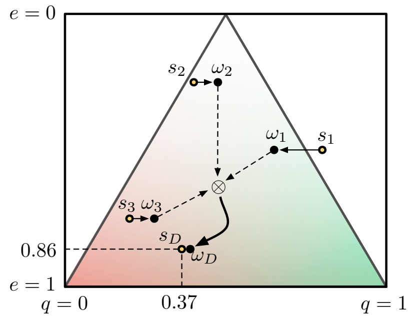

Figure 7: Combining three referee scores s_1, s_2, and s_3.

Let $scores(D) = \{s_1, \ldots, s_k\}$ be the set of scores for document D. Using the proposed transformation Δ, we can now compute the document's combined score s_D by

$$s_D = \Delta^{-1}(\Delta(s_1) \otimes \cdots \otimes \Delta(s_k)), \tag{14}$$

where Δ^{-1} denotes the inverse of Δ. Note that $\Delta^{-1}(\varphi) = (\frac{b}{1-i}, 1-i)$ is unique for all $\varphi = (b, d, i) \in \Phi$, except for the neutral opinion $(0, 0, 1)$. This is a mathematical imperfection, but it is of no importance for our particular application. Figure 7 shows three scores $s_1 = (0.8, 0.5)$, $s_2 = (0.4, 0.25)$, $s_3 = (0.2, 0.75)$, their combination $s_D = (0.37, 0.86)$, and corresponding opinions $\omega_i = \Delta(s_i)$ and $\omega_D = \omega_1 \otimes \omega_2 \otimes \omega_3 = \Delta(s_D)$.

3.2 Ranking Combined Referee Scores

Given the above solution for the score aggregation problem, we are now in possession of a set $\mathcal{S} = \{s_D : D \in \mathcal{D}\}$ of combined referee scores $s_D = (q_D, e_D)$, one for each document. The first thing to note here is the problem of documents with an unsatisfactory combined expertise level e_D. This means that the program committee or the editor was unable to assign the document to an appropriate referee. The usual procedure in such a case is to assign the paper to one or two additional referees. In our model, we may use a threshold $\gamma \in [0, 1]$ to define the set $\mathcal{D}_\gamma = \{D \in \mathcal{D} : e_D \leq \gamma\}$ of documents to be reassigned. The new referee scores are then combined with the existing ones to obtain an updated set of scores. This is a really important point when it comes to improve the quality of the review process, but time constraints usually do not allow more than one such iteration. Nierstrasz talks about the problem of low overall expertise levels in a pattern called *Identify Missing Champions* [26].

When all the updates are done and $\mathcal{S}$ is finally fixed, we reach the final stage of the review process, in which the decision about the accepted papers needs to be taken. An ideal basis for this decision would a be a ranking of the submitted documents, from which the highest-ranked submissions are accepted. We have seen in Section 3 that the set of Bayesian opinions is totally ordered, and that general opinions can be transformed into Bayesian opinions. The idea thus is to use the total order $\succeq_0$ of Bayesian opinions together with one of the proposed transformations to define an order with respect to Φ_z. More formally, we may consider three different orders $\succeq_g$, $\succeq_h$, and $\succeq_p$ for Φ, which are defined similarly by

$$\varphi \succeq_g \psi, \text{ iff } g(\varphi) \succeq_0 g(\psi), \quad \varphi \succeq_h \psi, \text{ iff } h(\varphi) \succeq_0 h(\psi), \text{ and } \varphi \succeq_p \psi, \text{ iff } p(\varphi) \succeq_0 p(\psi),$$

for all $\varphi, \psi \in \Phi_z$.[9] Note that $\succeq_g$, $\succeq_h$, and $\succeq_p$ are all *total preorders*, i.e. the antisymmetry property which is necessary for a total order does not hold. Nevertheless, we can use them to order the elements of $\mathcal{S}$, e.g. by

$$s_{D_1} \succeq_g s_{D_2}, \text{ iff } \Delta(s_{D_1}) \succeq_g \Delta(s_{D_2}), \text{ respectively } g(\Delta(s_{D_1})) \succeq_0 g(\Delta(s_{D_2})),$$

for g, and similarly for h and p. This is again a total preorder, which we can use to establish a document ranking for $\mathcal{D}$. Note that if $s_{D_1} \succeq_h s_{D_2}$ and $s_{D_2} \succeq_h s_{D_1}$ hold for two documents $D_1 \neq D_2$, which means that they have the same g-image in Φ_0, they are indistinguishable for $\succeq_g$ and thus receive the same rank. If necessary, such ties between equally ranked documents are broken at random.

[9] In the case of h, the order is only defined for Φ, i.e. it excludes the inconsistent opinion z obtained for two extreme opposite scores $(1, 1)$ and $(0, 1)$. To avoid this problem, either q should be restricted to $[0, 1)$, $(0, 1]$, or $(0, 1)$, or e should be restricted to $[0, 1)$. Another solution is to impose $h(z) = u$.

4 Conclusion

We have seen in this paper a solution for the score aggregation and document ranking problems. The key idea is to transform scores (q, e) into opinions (b, d, i), and to combine them by Dempster's rule. We have analyzed the underlying algebraic structures and properties. From various ways of projecting general opinions into the totally ordered set of Bayesian opinions, we can inherit the order and finally apply it to establish the final document ranking. The systematic formal analysis of this problem is the main contribution of this paper.

What is still missing today is the implementation of the proposed method in one of the existing conference management systems. A prototype implementation is available and can be tested at `http://www.iam.unibe.ch/~run/referee`, but it includes only the core of the proposed method with a very simple visualization. Nevertheless, it is interesting to observe that it almost perfectly reproduces some of Nierstrasz' classification patterns.

Another open issue is to apply and compare the recommended scheme empirically to real conference review data. It will certainly be interesting to observe whether real PC decisions are matched and to what extent. Note that we do not want to promote our method as a replacement for PC meetings or the discussions in editorial boards, but we think it could serve as a valuable decision support tool.

Acknowledgement

Thanks to Michael Wachter and Jacek Jonczy for careful proof-reading and to Milan Daniel and Oscar Nierstrasz for helpful discussions and comments. This research is partly supported by the SNF-project No. PP002–102652.

References

[1] Anonymous Author. A calculation of the credibility of human testimony. *Philosophical Transactions of the Royal Society*, 21:359–365, 1699.

[2] D. Black. *Theory of Committees and Elections*. Cambridge University Press, Cambridge, USA, 1958.

[3] G. Boole. *The Laws of Thought*. Walton and Maberley, London, 1854.

[4] O. Bottema. On the area of a triangle in barycentric coordinates. *Crux Mathematicorum*, 8:228–231, 1982.

[5] B. R. Cobb and P. P. Shenoy. On the plausibility transformation method for translating belief function models to probability models. *International Journal of Approximate Reasoning*, 41(3):314–330, 2006.

[6] F. Cuzzolin. Semantics of the relative belief of singletons. In *UncLog'08, International Workshop on Interval/Probabilistic Uncertainty and Non-Classical Logics*, number 46 in Advances in Soft Computing, pages 201–213, Ishikawa, Japan, 2008.

[7] M. Daniel. Algebraic structures related to the combination of belief functions. Technical Report 872, Academy of Sciences of the Czech Republic, Prague, Czech Republic, 2002.

[8] M. Daniel. On transformations of belief functions to probabilities. *International Journal of Intelligent Systems*, 21(3):261–282, 2006.

[9] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[10] A. P. Dempster. A generalization of Bayesian inference. *Journal of the Royal Statistical Society*, 30(2):205–247, 1968.

[11] J. R. Douceur. Paper rating vs. paper ranking. In *WOWCS'08, Workshop on Organizing Workshops, Conferences, and Symposia for Computer Systems*, 2008.

[12] M. Ginsberg. Non-monotonic reasoning using Dempster's rule. In *AAAI'84, 4th National Conference on Artificial Intelligence*, pages 112–119, Austin, USA, 1984.

[13] R. Haenni. Non-additive degrees of belief. In F. Huber and C. Schmidt-Petri, editors, *Degrees of Belief*. Springer, 2008.

[14] R. Haenni. Probabilistic argumentation. *Journal of Applied Logic*, 2008.

[15] R. Haenni and S. Hartmann. Modeling partially reliable information sources: a general approach based on Dempster-Shafer theory. *International Journal of Information Fusion*, 7(4):361–379, 2006.

[16] P. Hájek, T. Havránek, and R. Jiroušek. *Uncertain Information Processing in Expert Systems.* CRC Press, Boca Raton, USA, 1992.

[17] P. Hájek and J. J. Valdés. Generalized algebraic approach to uncertainty processing in rule-based expert systems (dempsteroids). *Computers and Artificial Intelligence*, 10:29–42, 1991.

[18] A. Jøsang. Artificial reasoning with subjective logic. In A. C. Nayak and M. Pagnucco, editors, *2nd Australian Workshop on Commonsense Reasoning*, Perth, Australia, 1997.

[19] A. Jøsang. A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(3):279–311, 2001.

[20] J. Kohlas. Reliability of arguments. In E. von Collani, editor, *Defining the Science of Stochastics*, Sigma Series in Stochastics, pages 73–94. Heldermann, Lemgo, Germany, 2004.

[21] J. Kohlas and P. A. Monney. *A Mathematical Theory of Hints – An Approach to the Dempster-Shafer Theory of Evidence*, Springer, 1995.

[22] E. Kohler. Hot crap! In *WOWCS'08, Workshop on Organizing Workshops, Conferences, and Symposia for Computer Systems*, San Francisco, USA, 2008.

[23] S. Krishnamurthi. The CONTINUE server (or, how i administered PADL 2002 and 2003). In *PADL'03, 5th International Symposium on Practical Aspects of Declarative Languages*, LNCS 2562, pages 2–16, New Orleans, USA, 2003.

[24] P. S. Laplace. *Théorie Analytique des Probabilités.* Courcier, Paris, 3ème edition, 1820.

[25] Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.* L'Imprimerie Royale, Paris, France, 1785.

[26] O. Nierstrasz. Identify the champion. In N. Harrison, B. Foote, and H. Rohnert, editors, *Pattern Languages of Program Design*, volume 4, pages 539–556. Addison-Wesley, 2000.

[27] M. Papagelis, D. Plexousakis, and P. Nikolaou. CONFIOUS: Managing the electronic submission and reviewing process of scientific conferences. In *WISE'05, 6rd International Conference on Web Information Systems Engineering*, LNCS 3806, pages 711–720, New York, USA, 2005.

[28] P. Rigaux. An iterative rating method: Application to web-based conference management. In *SAC'04, 19th Annual ACM Symposium on Applied Computing*, pages 1682–1687, 2004.

[29] G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, 1976.

[30] G. Shafer. The early development of mathematical probability. In I. Grattan-Guinness, editor, *Companion Encyclopedia of the History and Philosophy of the Mathematical Sciences*, pages 1293–1302, London, U.K., 1993. Routledge.

[31] P. Smets. Quantifying beliefs by belief functions: An axiomatic justification. In R. Bajcsy, editor, *IJCAI'93: 13th International Joint Conference on Artificial Intelligence*, pages 598–603, Chambéry, France, 1993.

[32] P. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–234, 1994.

[33] R. van de Stadt. CyberChair: a web-based groupware application to facilitate the paper reviewing process. available on-line at http://www.borbala.com/cyberchair/wbgafprp.pdf, 2001.

Rolf Haenni

Bern University of Applied Sciences
CH-2501 Biel, Switzerland
Email: `rolf.haenni@bfh.ch`

and

University of Bern
CH-3280 Bern, Switzerland
Email: `haenni@iam.unibe.ch`