A Formal Account of Effectiveness Evaluation and Ranking Fusion

Enrique Amigó
UNED NLP & IR Group
Madrid, Spain
enrique@lsi.uned.es

Fernando Giner
UNED NLP & IR Group
Madrid, Spain
fginer3@alumno.uned.es

Stefano Mizzaro
University of Udine
Udine, Italy
mizzaro@uniud.it

Damiano Spina
RMIT University
Melbourne, Australia
damiano.spina@rmit.edu.au

ABSTRACT

This paper proposes a theoretical framework which models the information provided by retrieval systems in terms of Information Theory. The proposed framework allows to formalize: (i) system effectiveness as an information theoretic similarity between system outputs and human assessments, and (ii) ranking fusion as an information quantity measure. As a result, the proposed effectiveness metric improves popular metrics in terms of formal constraints. In addition, our empirical experiments suggest that it captures quality aspects from traditional metrics, while the reverse is not true. Our work also advances the understanding of theoretical foundations of the empirically known phenomenon of effectiveness increase when combining retrieval system outputs in an unsupervised manner.

CCS CONCEPTS

• **Information systems → Information retrieval**;

KEYWORDS

information theory; evaluation; ranking fusion

ACM Reference Format:
Enrique Amigó, Fernando Giner, Stefano Mizzaro, and Damiano Spina. 2018. A Formal Account of Effectiveness Evaluation and Ranking Fusion. In *2018 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '18), September 14–17, 2018, Tianjin, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3234944.3234958

1 INTRODUCTION

Most of the research in the field of Information Retrieval (IR) is empirically-based. The effectiveness of retrieval approaches are typically validated over large data sets, most of them developed in the last decade. The effectiveness of ranking fusion and learning-to-rank algorithms are also validated in an empirical way. In addition, effectiveness metrics are supported by empirical user behavior studies or meta-metrics such as robustness or sensitivity [3, 17].

On the other hand, some works aim to provide explanations for some phenomena observed in empirical experiments. For instance, the Probability Ranking Principle [34] assumes that retrieval systems return documents ranked in order of decreasing probability of relevance to the user. In the same way, based on empirical user behavior observations, evaluation metrics are supported by the top-heaviness principle [9], which gives more weight to highly-ranked documents in the evaluation process. Studies in unsupervised ranking fusion algorithms have reported empirically that the most effective combinations of rankings are those in which the relevant documents are unanimously early-ranked, while the retrieved non-relevant documents vary across rankings [24, 35]. Likewise, other studies have reported empirically that human assessments can be replaced successfully –at least to some extent– by the average of system outputs in an evaluation campaign [6, 33].

In this paper, we aim to define a theoretical framework which models the phenomena described above. The framework is based on the notion of *Observational Information Quantity*: Rather than focusing on document content, this framework models the information provided by retrieval systems (document rankings) and human assessors in terms of Information Theory [31]. On the basis of observational information quantity we then define an entropy-like notion that allows the formalization of system effectiveness as an information-theoretic similarity between system outputs and human assessments. The proposed framework also models ranking fusion as an information quantity measure.

The resulting effectiveness metric improves most of existing metrics in terms of formal constraints. In other words, the proposed framework gives a basis –grounded in Information Theory– for effectiveness metrics, which were traditionally supported by user behavior modeling. Additionally, our experiments corroborate this analysis, showing that the proposed metric captures quality aspects from traditional metrics, while the reverse is not true.

On the other hand, our work provides a theoretical foundations of the empirically-known phenomenon of effectiveness increase when combining retrieval system outputs in an unsupervised manner. Our experiments also check empirically the assumptions in which the proposed theoretical framework is grounded.

Let us remark that this work does not attempt to provide *better* solutions than those presented in previous work; rather we aim at

ICTIR '18, September 14–17, 2018, Tianjin, China

ACM ISBN 978-1-4503-5656-5/18/09…$15.00
https://doi.org/10.1145/3234944.3234958

defining a global theoretical framework on which to base future improvements.

The rest of the paper is organized as follows. Section 2 discusses related work and Section 3 introduces the theoretical framework based on Observational Information Quantity. Section 4 analyzes how our proposed framework can be used to inform an effectiveness evaluation metric that satisfies a set of formal constraints. Section 5 describes the justification of ranking fusion based on Observational Information Quantity. Section 6 connects the definitions of our framework with those of the classical Information Theory by Shannon. Finally, Section 7 concludes the work.

2 RELATED WORK

2.1 Measuring Effectiveness

Most of current metrics estimate effectiveness by assuming an underlying user model for browsing relevant and non-relevant documents returned in the system output ranking. For instance, Discount Cumulative Gain [20] assumes that the probability of exploring deeper ranking positions decreases in a logarithmic manner. Expected Reciprocal Rank (ERR) [10] assumes a cascade model in which the user is looking for a particular document. Rank-Biased Precision (RBP) [27] assumes that a fixed probability of exploring the next document in the ranking. According to the analysis by Amigó et al. [3] none of the most popular metrics satisfies completely a set of five formal constraints. RBP satisfies four of them but not the *confidence* constraint, which penalizes the addition of non-relevant documents at the end of the ranking.

Some authors have focused on explaining evaluation metrics in terms of Measurement Theory, tackling the issue of the suitability of the interval scale assumption [16] or interpreting metrics as an homomorphism (measurement) between effectiveness and systems [15]. These works state formal constraints and desirable properties, but they do not derive any particular approach.

In this paper, we apply an information theory-based similarity measure to compare system outputs against the gold-standard. Our theoretical analysis shows that user behavior based constraints can be satisfied by grounding the metric in information theory principles.

2.2 Ranking Fusion

Finding a theoretical explanation for the effectiveness of combining system outputs in an unsupervised manner has been largely explored in the literature. This problem has been modeled from two closely-related perspectives: classifier ensembles and ranking fusion.

From the first perspective, the literature shows that combining classifiers is effective when the individual classifiers are accurate and diverse. Hansen and Salamon [18] proved that if the average error rate for an example is less than 50% and the component classifiers in the ensemble are independent in the production of their errors, the expected error for that example can be reduced to zero as the number of classifiers combined goes to infinity. This theoretical analysis was actually reported by the Condorcet's jury theorem in 1785 [7, 12]. However, such assumptions rarely hold in practice. Krogh and Vedelsby [22] later formally showed that an ideal ensemble consists of highly correct classifiers that disagree as much as possible. In general, a point of consensus is that when the classifiers make statistically independent errors, the combination has the potential to increase the performance of the system. Other studies assume correlation between signals, but equal performance and homogeneous correlation [21], which is also non-realistic in the context of information systems. Matan [26] analyzed the upper and lower bounds of classification of a majority based ensemble. In the particular context of information retrieval tasks, Shaw et al. [32] found that the best combination strategy consisted of summing the outputs of the retrieval algorithms, and Hull et al. [19] found that the best improvement in performance in the context of a filtering task came from the simple averaging strategy.

Table 1: Example of system outputs and human assessments.

Rank	Γ			
	r_1	r_2	r_3	Human assessments g
1	$r_1(d_1)$	$r_2(d_3)$	$r_3(d_3)$	$g(d_1) = g(d_4) = 1$
2	$r_1(d_2)$	$r_2(d_1)$	$r_3(d_1)$	
3	$r_1(d_4)$	$r_2(d_2)$	$r_3(d_2)$	$g(d_{i=2,3,5,6,\ldots}) = 0$
n/a (∞)	$r_{1,2,3}(d_5, d_6, \ldots) = 0,$		$r_{2,3}(d_4) = 0$	

From the ranking fusion perspective, Montague and Aslam [28] reported an important improvement of unsupervised combined systems w.r.t. the best single system in multiple TREC test beds. In addition, just like in the classification scenario, the need for avoiding redundant systems has been reported in the context of ranking fusion. For instance, Nuray-Turan and Can [29] reported effectiveness improvement when selecting rankings that differ from the majority voting in the ranking fusion process. Lee [24] and later Vogt and Cottrell [35] found that the best combinations were between systems that retrieve similar sets of relevant documents and dissimilar sets of non-relevant documents. There exist other works that reformulate ranking fusion algorithms in terms of probability estimations, always under the independence assumption [5, 8, 25]. Finally, Amigó et al. [1] proposed an extension of the notion of Information Quantity in order to generalize different ranking fusion methods in the context of text similarity. In this paper, we review this notion extending it to *observational entropy*.

3 OBSERVATIONAL INFORMATION QUANTITY

3.1 An Example

Let us start with a simple example that considers the output of a set of information retrieval systems as in Table 1. The collection $\mathcal{D}$ contains a large amount of documents. There are three systems s_1, s_2, and s_3 that return different documents producing rankings r_1, r_2, and r_3 of length 3. We assume that documents out of these rankings share the same infinite rank.

Rather than considering the content of the documents, we *observe* the documents from the perspective of a set of information retrieval systems. The first step should be being able to *measure* the quantity of information provided by systems or human assessments for single documents. This information measurement should satisfy the following three properties.

First, the more a document is highly ranked according to a retrieval system, or relevance scored according to human assessors, the more the document is discriminated against the large collection (increasing informativeness). For instance, according to r_1 in the example, we have more information about the relevance of d_1 than d_2.

Second, the earlier documents are ranked according to different systems, the more information about their relevance we observe. This is in line with the conclusions found in previous work [24, 35]: the most effective combinations of rankings are those in which the relevant documents are unanimously early ranked, while the retrieved non-relevant documents vary across rankings. For instance, d_1 occurs at the first or second position in every ranking. Therefore, we have more information to estimate the relevance of d_1 than other documents.

Third, redundant systems provide less information than non-redundant systems. In relation to this, the profits of combining non-redundant systems have been reported in both the ranking [29] and classification [22] scenarios. For instance, d_3 is ranked in the first position by r_2 and r_3, but both rankings seems to be similar, that is, they seem to be providing the same information.

3.2 The Framework

Let us consider a set of *relevance signals*, Γ, which consists of the set of rankings and human assessments, $\Gamma = \{r_1, \ldots, r_n, g\}$[1], which assign scores to documents in a collection of documents, $\mathcal{D}$. Then, we define *unanimous outscoring* as follows.[2]

Definition 3.1. A document, d, is unanimously outscored by another document, d', according to a set of signals, Γ, whenever it is outscored for every signal simultaneously:

$$d' \geq_\Gamma d \iff \forall \gamma \in \Gamma.\, \gamma(d') \geq \gamma(d).$$

In the following we will use both $\leq_\Gamma$ and $\geq_\Gamma$, with the obvious meaning. Going back to the example in Table 1, we obtain the following outscoring relationships regarding to $d_1, \ldots, d_4$. Being $\Gamma = \{r_1, r_2, r_3, g\}$ (note that we also take into account the human assessments in the ground truth g):

$$\begin{aligned} &d_1 \geq_\Gamma d_1 \quad d_1, d_2 \geq_\Gamma d_2 \\ &d_3 \geq_\Gamma d_3 \quad d_1, d_4 \geq_\Gamma d_4. \end{aligned} \tag{1}$$

Documents d_1 and d_3 are only outscored by themselves, i.e., there is no other documents that is unanimously ranked earlier in Γ. d_2 is outscored by d_2 and d_1 (including g). Likewise, d_4 is outscored by itself and also by d_1, given that it is corroborated by the three rankings and the gold g.

Then, the observational information quantity of a document is defined as follows.

Definition 3.2. The Observational Information Quantity, $\mathcal{I}_\Gamma\left(d\right)$, of a document, d, under a set of signals, Γ, is the minus logarithm of the probability of being unanimously outscored by other documents:

$$\mathcal{I}_\Gamma\left(d\right) = -log\left(P_{d' \in \mathcal{D}}\left(d' \geq_\Gamma d\right)\right). \tag{2}$$

In other words, the more a document is unanimously outscored simultaneously in all signals by other documents, the less the document is informative. In consequence, highly informative documents are those that are highly scored by all rankings and the human assessment.

For instance, going back to the example in Table 1 and taking into account inequalities in Eq. (1), the Observational Information Quantity of document documents is:[3]

$$\mathcal{I}_\Gamma\left(d_1\right) = -\log\left(\frac{1}{|\mathcal{D}|}\right), \qquad \mathcal{I}_\Gamma\left(d_2\right) = -\log\left(\frac{2}{|\mathcal{D}|}\right),$$
$$\mathcal{I}_\Gamma\left(d_3\right) = -\log\left(\frac{1}{|\mathcal{D}|}\right), \qquad \mathcal{I}_\Gamma\left(d_4\right) = -\log\left(\frac{2}{|\mathcal{D}|}\right).$$

Documents that obtain the lowest score by all signals (i.e., documents that are both non-retrieved and non-relevant) obtain the lowest Observational Information Quantity, as they are outscored by d_1, d_2, d_3, d_4 and by themselves:

$$\forall i \in \{5, 6, \ldots\}.\, \mathcal{I}_\Gamma\left(d_i\right) = -\log\left(\frac{|\mathcal{D}|}{|\mathcal{D}|}\right) = 0.$$

Our formalization of observational information quantity matches with the definition provided by Amigó et al. [1] for similarity measures fusion. Here, we extend it to define an entropy-like notion.

Definition 3.3. The Observational Entropy, $H\left(\Gamma\right)$, of a set of signals, Γ, is the expected observational information quantity across the document set, $\mathcal{D}$:

$$H\left(\Gamma\right) = \frac{\sum_{d \in \mathcal{D}} \mathcal{I}_\Gamma\left(d\right)}{|\mathcal{D}|}.$$

Intuitively, the observational entropy of a set of signals represents the extent to which finding unanimous improvement is unlikely. Thus, non-correlated signals will tend to achieve a higher entropy. Also note that this definition is inspired by the classical Shannon's Entropy, but it differs from the original because it uses two different probability distributions: the probability distribution of outscoring (the one used in Equation (2)) and the probability distributions of a document ($\frac{1}{|\mathcal{D}|}$). Section 6 explains the connection between traditional and observational information quantity. For the sake of simplicity, we will denote hereafter the entropy for a single signal set as $H\left(\gamma\right)$.

3.3 Properties

Observational Information Quantity and Observational Entropy satisfy the following general properties, that will be useful in the following. [4]

Property 3.1. *For all $\gamma \in \Gamma$, $\mathcal{I}_\Gamma\left(d\right)$ is monotonic w.r.t. signal values $\gamma(d)$:*

$$d_1 \geq_\Gamma d_2 \Longrightarrow \mathcal{I}_\Gamma\left(d_1\right) \geq \mathcal{I}_\Gamma\left(d_2\right).$$

Property 3.1 implies the following corollary.

Corollary 3.2. *The observational information quantity of a document under a single signal grows with its signal value:*

$$\mathcal{I}_{\{\gamma\}}\left(d\right) \propto \gamma(d).$$

[1] We will use $\gamma \in \Gamma$ to refer to signals in a general manner.
[2] We assume that earlier rank positions correspond with higher scores.

[3] Note that the probabilities are computed as frequencies.
[4] See formal proofs in Appendix A.

PROPERTY 3.3. *Both observational entropy and observational information quantity do not decrease when adding signals to the set* Γ. *Given a signal* $\gamma \notin \Gamma$*:*

$$\mathcal{I}_{\Gamma\cup\{\gamma\}}\Big(d\Big) \geq \mathcal{I}_{\Gamma}\Big(d\Big)$$
$$H\Big(\Gamma \cup \{\gamma\}\Big) \geq H\Big(\Gamma\Big).$$

PROPERTY 3.4. *The observational entropy of a single ranking depends exclusively on its length. More formally, being* γ *a signal such that*

$$\gamma(d_1) > \gamma(d_2) > \ldots > \gamma(d_n) > \gamma(d_{n+1}) = \gamma(d_{n+2}) = \ldots,$$

then:

$$H\Big(\gamma\Big) = -\frac{1}{|\mathcal{D}|}\sum_{i=1}^{n} log\left(\frac{i}{|\mathcal{D}|}\right).$$

PROPERTY 3.5. *Observational entropy and observational information quantity are invariant under redundant signals. Being* f *any strict monotonic function (i.e., a function that does not affect the ordinal relationships)*

$$\mathcal{I}_{\Gamma\cup\{\gamma\}}\Big(d\Big) = \mathcal{I}_{\Gamma\cup\{\gamma,f(\gamma)\}}\Big(d\Big)$$
$$H\Big(\Gamma \cup \{\gamma\}\Big) = H\Big(\Gamma \cup \{\gamma, f(\gamma)\}\Big).$$

This aspect is crucial when representing documents in terms of systems output signals. Redundant systems should not increase the observational information quantity of documents. In addition, adding a non-redundant signal increases strictly the entropy of the signal set.

PROPERTY 3.6. *If a preference between two documents in* γ *is not corroborated by any signal in* Γ*, i.e.,*

$$\exists d_1, d_2 \in \mathcal{D}. \left(d_1 \geq_{\{\gamma\}} d_2 \wedge d_1 \leq_{\Gamma} d_2\right),$$

then the entropy strictly increases when adding the signal to the set, i.e.,

$$H\Big(\Gamma\Big) < H\Big(\Gamma \cup \{\gamma\}\Big).$$

4 MEASURING EFFECTIVENESS

We now show how the above definitions and properties can be exploited to define an effectiveness measure that satisfies formal constraints which are not satisfied by traditional metrics.

4.1 Observational Information Effectiveness

As we said in the related work section, instead of modeling the user behavior in the seeking process, we apply an information theory based similarity measure to compare system outputs against the gold-standard, but using our notions of observational information quantity and entropy. More specifically, we use the Information Contrast Model (ICM) [2]. ICM is a parameterizable extension of Point-wise Mutual Information which satisfies a set of constraints whenever its parameter are within a certain range [2]. We start by defining a notion of effectiveness for a signal.

Definition 4.1. Given a signal, γ*, and a relevance gold standard,* g*, then the Observational Information Effectiveness (OIE) of the signal is a linear combination of observational entropies as follows:*

$$\mathrm{OIE}(\gamma, g) = \alpha_1 \cdot H\Big(\gamma\Big) + \alpha_2 \cdot H\Big(g\Big) - \beta \cdot H\Big(\{\gamma, g\}\Big). \quad (3)$$

where $\alpha_1, \alpha_2, \beta \in \mathbb{R}^+$ and α_1 and α_2 weight the effect of the ranking and assessment entropy. Hereafter, we will consider $\alpha_1 = \alpha_2 = 1$ for the sake of simplicity. Note that the entropy of a single ranking with a fixed length is constant (Property 3.4). Therefore, when evaluating single rankings (i.e., when $\Gamma = \{r\}$) with a fixed length under a fixed gold, the parameters α_1 and α_2 do not affect the relative effectiveness of systems, which depends exclusively on the component $H\Big(\{\gamma, g\}\Big)$.

4.2 Satisfying Formal Constraints

We are interested in comparing the proposed metric with the state-of-the-art. Comparing metrics empirically against user satisfaction or search effectiveness requires data that is often unavailable and expensive to collect. Moreover, findings may be biased to the subjects, retrieval systems or other experimental factors.

An alternative consists of studying evaluation metrics under formal constraints. Amigó et al. [3] defined a theoretical framework according to five formal constraints: swapping contiguous documents in concordance with the gold increases effectiveness (*priority constraint*, `Pri`); the effect of swapping is larger at the top of the ranking (*deepness constraint*, `Deep`); retrieving one relevant document is better than a huge amount of relevant documents after a huge set of irrelevant documents (*deepness threshold constraint*, `DeepTh`); there exists a certain area at the top of the ranking in which n relevant documents is better than only one (*closeness threshold constraints*, `CloseTh`); and finally, adding irrelevant documents at the bottom of the ranking decreases effectiveness (*confidence constraint*, `Conf`). According to this study, among the most popular metrics, only the Rank-Biased Precision (RBP) metric [27] satisfies the first four constraints. The following theorem states that OIE satisfies these five constraints.

THEOREM 4.2. **Information Evaluation Theorem** *OIE satisfies the five constraints defined by Amigó et al. [3] whenever* $1 < \beta < \frac{2n-1}{n}$*, being n the minimum amount of documents that are necessarily explored by the user.*[5]

Surprisingly, these theoretical boundaries for β correspond with those predicted by Amigó et al. [2] for the ICM similarity model, even though ICM is grounded on a different axiomatics, oriented to the general notion of similarity.

4.3 Experiment

Although this work is mainly theoretical, we performed a brief experiment comparing OIE against traditional metrics. Here, we use the meta-metric *Metric Unanimity (MU)* [4]. MU quantifies to what extent a metric is sensitive to quality aspects captured by other existing metrics. The intuition is that, if a system improves another system for every quality criteria, this should be *unanimously* reflected by every metric. A metric that captures every quality criteria should reflect these improvements.

MU is formalized as the Point-wise Mutual Information (PMI) between metric decisions and improvements corroborated by all the metrics in a given set of metrics, $\mathcal{M}$. Formally, given a metric,

[5] n is a variable defined for the closeness deepness threshold constraint [3].

Table 2: Traditional metrics and Observational Information Effectiveness (OIE), ranked by Metric Unanimity (MU) [4]. ● indicates that the metric satisfies the formal constraint, ○ indicates otherwise.

Metric	MU	Pri	Deep	DeepTh	CloseTh	Conf
$\mathrm{OIE}_{\beta=1.2}$	0.928	●	●	●	●	●
$\mathrm{OIE}_{\beta=1}$	0.927	●	●	●	○	○
RBP	0.926	●	●	●	●	○
DCG	0.914	●	●	○	●	○
AP	0.910	●	●	○	●	○
P@100	0.910	○	○	●	●	○
DCG@50	0.905	○	○	●	●	○
ERR@50	0.903	○	○	●	○	○
ERR	0.901	●	●	●	○	○
P@50	0.900	○	○	●	●	○
ERR@20	0.886	○	○	●	○	○
DCG@20	0.886	○	○	●	●	○
P@20	0.876	○	○	●	●	○
P@10	0.829	○	○	●	●	○
RR@10	0.162	○	○	●	○	○

$m \in \mathcal{M}$, and a set of system outputs, $\mathcal{R}$:[6]

$$\mathrm{MU}_{\mathcal{M},\mathcal{R}}(m) = \mathrm{PMI}\left(\Delta m_{i,j}, \Delta\mathcal{M}_{i,j}\right) = \log\left(\frac{P(\Delta m_{i,j}, \Delta\mathcal{M}_{i,j})}{P(\Delta m_{i,j})\cdot P(\Delta\mathcal{M}_{i,j})}\right).$$

In the equation, $\Delta m_{i,j}$ and $\Delta\mathcal{M}_{i,j}$ are statistical variables over system pairs $(r_i, r_j) \in \mathcal{R}^2$, indicating a system improvement according to the metric and to every metric, respectively:

$$\Delta m_{i,j} \equiv m(r_i) > m(r_j)$$
$$\Delta\mathcal{M}_{i,j} \equiv \forall m \in \mathcal{M}\left(m(r_i) \geq m(r_j)\right).$$

Note that MU is closely related with the unanimity notion in observational information quantity. The reader could thing that there exists some theoretical over-fitting here. However, MU and observational information quantity ($\mathcal{I}_\Gamma\left(d\right)$) measure different things. $\mathrm{MU}_{\mathcal{M},\mathcal{R}}(m)$ measures the correspondence between a metric and the whole set of metrics (system rankings), while $\mathcal{I}_\Gamma\left(d\right)$ combines system outputs (document rankings).

In all our experiments, we used the Gov-2 collection and the topics 701 to 750 used in the TREC 2004 Terabyte Track [11]. We also used the 60 official runs submitted by the participants to the track. Table 2 shows the MU results for OIE with $\beta = 1.2$ (an arbitrarily selected value in the theoretical grounded range), and other standard evaluation metrics including: OIE with $\beta = 1$ (out of the theoretically grounded range), Precision at cutoff k (P@k), Average Precision (AP), Reciprocal Rank (RR) [36], Expected Reciprocal Rank (ERR@k) [10], Discounted Cumulative Gain (DCG@k) [20] and Rank-Biased Precision (RBP) [27]. For OIE, we have considered the cutoff point at position 100. For the rest of metrics, we have considered the cutoff positions 20, 50 and 100. For the RBP metric, we have considered the values 0.8, 0.9 and 0.99 for the parameter p.

As the table shows, metrics with shallow cutoffs (20 or 50) and RR, which stops at the first relevant document, are at the bottom of the MU ranking, given that they capture only partial aspects of the ranking quality. On the other hand, $\mathrm{OIE}_{\beta=1.2}$ improves the rest of metrics in terms of MU. The improvement of $\mathrm{OIE}_{\beta=1.2}$ regarding $\mathrm{OIE}_{\beta=1}$ corroborates the theoretical analysis about the β ranges.

Interestingly, RBP (the third ranked metric) is the only one that satisfies the four first constraints according to the study by Amigó et al. [3]. Note that we obtained the same Unanimity for RBP regardless the p parameter ($p \in \{0.8, 0.9, 0.99\}$). The improvement in MU for OIE compared against RBP is probably due to the *Confidence* constraint. Some rankings have less than 100 documents. The benefit of avoiding non-relevant documents at the end of the ranking is only rewarded by OIE in this metric set. Note also that some metrics such as DCG satisfy the Deepness Threshold constraint when adding a ranking cutoff, but this is at the cost of Priority and Deepness, given that documents at deep positions are not considered.

This experiment can be extended for more metrics, meta-evaluation criteria, and data sets. In this paper we focus on the Observational Information Quantity as a theoretical framework that can be applied to different phenomena in IR. Our formal analysis shows that effectiveness can be grounded in the Observational Information framework.

5 UNSUPERVISED RANKING FUSION

We now turn to applying the framework to the problem of ranking fusion.

5.1 Information Quantity Cumulative Evidence

As we mentioned in the introduction, the experience in most of experiments reported in the IR literature corroborates the Probability Ranking Principle. For instance, precision/recall curves tend to be descendant for every system. Given that $\mathcal{I}_r\left(d\right)$ is correlated with the position of d in the ranking r (Property 3.2), we can say that the probability of relevance increases with the Observational Information Quantity under a single signal. We generalize this phenomenon for multiple signals stating the following assumption.

Assumption 1. *[Information Quantity Cumulative Evidence] Adding signals increases the probability of improving relevance under an Observational Information Quantity increase*

$$P\left(d \geq_g d_0 \,\middle|\, d \geq_{\mathcal{I}_{\Gamma\cup\{\gamma\}}} d_0\right) \geq P\left(d \geq_g d_0 \,\middle|\, d \geq_{\mathcal{I}_\Gamma} d_0\right).$$

Note that according to Property 3.2, the observational information quantity $\mathcal{I}_\gamma$ under a single signal γ shorts documents in the same manner than γ. Therefore, we can directly infer from this theorem that an increase according to the $\mathcal{I}_\Gamma$ is more reliable than an increase of signals in isolation.

In order to check this assumption empirically, we make a pooling from the first 100 ranked document for each system output in our data set. Here, we consider that documents not present in a particular ranking γ are scored with the lowest signal value according to γ. In each experiment:

(i) we randomly select one topic and a set Γ of five system outputs;

[6] Note that the a priori probability of system improvement for every metric is fixed $P(\Delta m_{i,j}) = \frac{1}{2}$. That is, in the cases where two system outputs obtain the same score $m(s_i) = m(s_j)$, we add 0.5 to the statistical count.

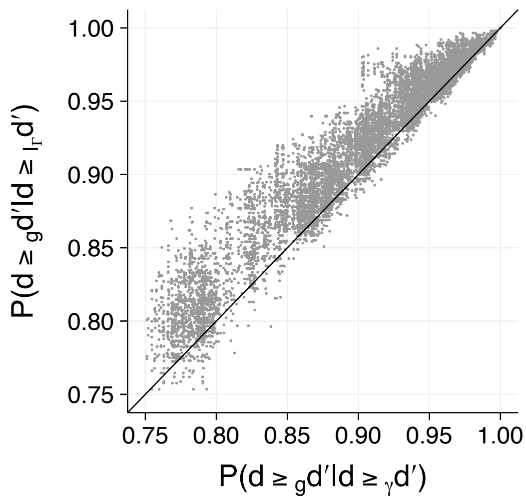

Figure 1: Checking the Information Effectiveness Additivity proposition.

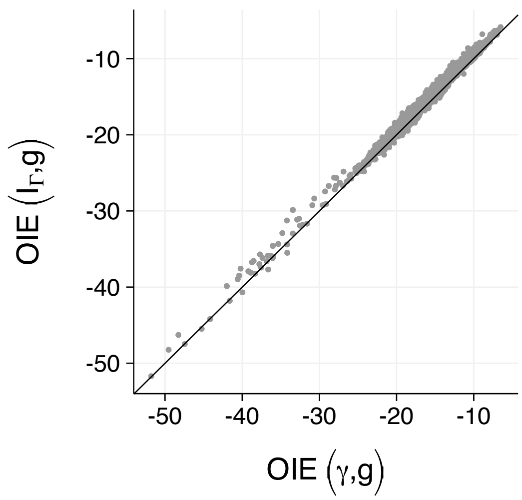

Figure 2: Experiment for checking the optimality of $\mathcal{I}_\Gamma$ as ranking fusion method.

(ii) we compute the observational information quantity $\mathcal{I}_\Gamma\left(d\right)$ for each document under this set of measurements Γ;
(iii) we select one single signal (one system output) γ from Γ;
(iv) We compute the conditional probabilities $P\left(d \geq_g d'|\gamma(d) \geq \gamma(d')\right)$ and $P\left(d \geq_g d'|d \geq_{\mathcal{I}_\Gamma} d'\right)$. That is, the relevance increase, when increasing the system score according to γ and when increasing the observational information quantity $\mathcal{I}_\Gamma$.

Each dot in Figure 1 represents one experiment, thus, one topic, five signals and one single signal from this set. The horizontal axis represents $P\left(d \geq_g d'|d \geq_\gamma d'\right)$. The vertical axis represents $P\left(d \geq_g d'|d \geq_{\mathcal{I}_\Gamma} d'\right)$. As the figure shows, the Information Effectiveness Additivity proposition practically always holds.

5.2 Ranking Fusion by Observational Information Quantity

The ranking fusion model proposed in this paper consists of the combination of system outputs in a single signal according to the observational information quantity of documents. That is $\gamma_{\text{Fusion}}(d) = \mathcal{I}_\Gamma(d)$. Then, we can state the following theorem.

THEOREM 5.1. *[Ranking Mergeability:] Under the Information Quantity Cumulative Evidence, and assuming that the information quantity estimation is fine grained, the effectiveness of $\mathcal{I}_\Gamma$ for any β values in the interval $(1,2)$ is higher than the effectiveness of any single measurement $\gamma \in \Gamma$:*

$$OIE_{\beta\in(1,2)}\left(\mathcal{I}_{\Gamma\cup\{\gamma\}},g\right) \geq OIE_{\beta\in(1,2)}\left(\mathcal{I}_\Gamma,g\right).$$

This theorem has strong practical implications. For instance, it means that instead of evaluating five systems, we can directly join them to achieve the best result. However, the Effectiveness Additivity Theorem has an important limitation, which is the need for high granularity in $\mathcal{I}_\Gamma$. That is, we need a huge amount of documents and an extremely costly computation process to accurately estimate the probability of improvement for all measurements. In addition, there exist some theoretical limitations for the granularity in particular situations. For instance, two documents appearing at the first position of two different rankings have necessarily the same $\mathcal{I}_\Gamma$, which is $\frac{1}{|\mathcal{D}|}$.

Let us check the Ranking Mergeability theorem empirically. To this aim, we emulate fine-grained single signals and $\mathcal{I}_\Gamma$ as follows. We start by generating random samples of five signals Γ, and one single signal γ from Γ. Then, we collect documents from the single ranking γ progressively from the top to the bottom, but discarding documents that achieve the same $\mathcal{I}_\Gamma$ than previously collected documents. This will generate a set of documents $\mathcal{D}'$ such that:

$$\forall d_1, d_2 \in \mathcal{D}'.\, \mathcal{I}_\Gamma\left(d_1\right) \neq \mathcal{I}_\Gamma\left(d_2\right) \wedge \forall\gamma \in \Gamma.\, \gamma(d_1) \neq \gamma(d_2).$$

Finally, we compare the effectiveness of γ and $\mathcal{I}_\Gamma$ in terms of OIE (using $\beta = 1.2$) across all topics, without considering the rest of documents.

Figure 2 compares the effectiveness of the single signal OIE(γ, g) (x-axis) against the effectiveness of the combined signals OIE$(\mathcal{I}_\Gamma, g)$ (y-axis). In our experiments, OIE$(\mathcal{I}_\Gamma, g) >$ OIE(γ, g) for $1,809$ out of $2,000$ cases.

As the theory predicts, $\mathcal{I}_\Gamma$ outperforms the single signal in almost all the cases. Notice that the improvements are less prominent than in the previous experiment. The first reason is that we are considering rankings instead of probabilities under non strict comparisons between signal values. The second reason is that OIE captures *top heaviness*, giving more weight to documents located at the top of rankings.

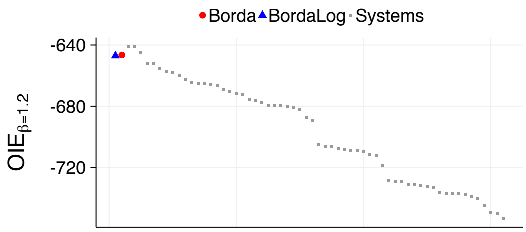

Figure 3: Evaluating the Borda fusion and the logarithmic Borda (BordaLog) over all system outputs.

5.3 Borda Count as an Approach to Observational Information Quantity

An important drawback of the Observational Information Quantity as a ranking fusion method is the need for a huge amount of samples in the estimation process. The reason is that the probability of unanimous outscoring decreases dramatically when adding signals. A common way of estimating joint probabilities under a limited amount of data consists of assuming independence across variables ($P(A,B,C) \simeq P(A) \cdot P(B) \cdot P(C)$). In this subsection we show that $\mathcal{I}_\Gamma$ converges into a Borda Count variant consisting of averaging the logarithm of rank positions instead of the original ranking position. Note that Borda Count is the most popular way for averaging rankings. As described in Section 3, our $\mathcal{I}_\Gamma$ definition matches with the Information Quantity defined by Amigó et al. [1] for similarity measure fusion. In their paper it is proved that this definition converges to other common ranking fusion algorithms depending on statistical assumptions. In particular, they prove that:

Proposition 5.2. *Assuming statistical independence, the Observational Information Quantity of a document under a set of rankings corresponds with the average logarithm of ranking positions:*

$$\mathcal{I}_\Gamma\left(d\right) \propto Avg_{\gamma_i \in \Gamma} \log(rank_i(d)).$$

Therefore, according to our analysis, in terms of Observational Information Effectiveness, if the independence assumption can be assumed, the $\text{Borda}_{\log}$ ranking fusion should be at least as effective than the best single system output. In addition, $\text{Borda}_{\log}$ should achieve a similar or better effectiveness than the original Borda algorithm.

Figure 3 illustrates the effectiveness of the original Borda, $\mathcal{I}_\Gamma$ under the independence assumption (BordaLog) and each system in our experimental data set. We have used the β value 1.2 according to the experiments in previous section. In order to evaluate outputs under the same conditions (fixed ranking length), we have truncated the Borda and BordaLog outputs in position 100 just like single systems. The figure shows that, in concordance with the theoretical analysis, both fusion methods practically achieve the same performance as the best system in the combination.

6 CONNECTING WITH TRADITIONAL INFORMATION THEORY

Probably, there are many possible theoretical explanations for the observational information framework. Signals (retrieval system outputs) in IR are quantitative, while the traditional information theory measures the information quantity of events characterized by binary features. On the other hand, *Differential Entropy* considers a continuous space of signals, but it does not allow to estimate the information quantity of a single event in this continuous space.

We now describe our proposed derivation for the observational information framework. We start by representing object observations as fuzzy sets. This allows us to capture both the amount of signals and their quantitative projection into each object. Then, we use the Dempster-Shafer theory of evidence [14, 30]. In particular, we use the belief function over the fuzzy set operators in order to estimate the information quantity of observations. Note that the proposed model differs from other approaches based on Demster-Shafer theory which focus on document content representation [13, 23].

A fuzzy set is formally defined as:

Definition 6.1. A fuzzy set is a pair (A, f) where A is a set and f a membership function $f : A \longrightarrow [0,1]$.

Then, an observation can be formalized as follows.

Definition 6.2. Given a set of signals, $\Gamma = \{\gamma_1, \ldots, \gamma_n\}$, and a set of possible values generated by each signal, $\{x_1, \ldots, x_n\}$, an observation, $\mathcal{O}_\Gamma(x_1, \ldots, x_n)$, under Γ is a fuzzy set of signals whose membership function corresponds to the signal values: $A = \Gamma$ and $f(\gamma_i) = x_i, \forall i \in \{1, \ldots, n\}$.

In other words, a document observation has two main components: the signals under which the document is observed and the corresponding signal values. From the previous definition we can infer that each document, $d \in \mathcal{D}$, produces an observation, $\mathcal{O}_\Gamma(\gamma_1(d), \ldots, \gamma_n(d))$, denoted as $\mathcal{O}_\Gamma(d)$.

According to the inclusion operator in fuzzy sets, an observation is included into another when:

$$\mathcal{O}_\Gamma(x_1, \ldots, x_n) \subseteq \mathcal{O}_\Gamma(x'_1, \ldots, x'_n) \iff \forall i = 1, \ldots, n; x'_i \geq x_i.$$

The purpose of Dempster-Shafer's theory is to represent believes in a set of elements referred to as a *frame of discernment*. We can consider the *believe function* defined on the observations of documents, this is possible by taking into account the inclusion relationship between observations.

Then, applying the Dempster-Shafer evidence theory, we define the mass function or *Basic Probability Assignment (BPA)* as the probability of observations across the set of documents $\mathcal{D}$. Being ω an observation:

$$m(\omega) = P_{d \in \mathcal{D}}\left(\mathcal{O}_\Gamma(d) = \omega\right).$$

Consequently, the corresponding *belief function* of an observation ω' is:

$$Bel(\omega') = \sum_{\omega | \omega \supseteq \omega'} m(\omega).$$

And therefore, the belief of a document observation can be expressed as:

$$Bel(\mathcal{O}_\Gamma(d)) = \sum_{\omega | \omega \supseteq \mathcal{O}_\Gamma(d)} P_{d' \in \mathcal{D}}\left(\omega = \mathcal{O}_\Gamma(d')\right) = P_{d' \in \mathcal{D}}\left(\mathcal{O}_\Gamma(d') \supseteq \mathcal{O}_\Gamma(d)\right).$$

The last identity is attained by considering that observations are actually a partition of the document space.

Then, the observational information quantity is analogous to the information quantity in Shannon's theory but replacing the probability with the belief function:

$$\begin{aligned} I_\Gamma(d) = I(O_\Gamma(d)) &= -\log(Bel(O_\Gamma(d))) \\ &= -\log(P_{d' \in \mathcal{D}}(O_\Gamma(d') \supseteq O_\Gamma(d))) \\ &= -\log(P_{d' \in \mathcal{D}}(d' \geq_\Gamma d)), \end{aligned}$$

which leads directly to Definition 3.2.

The entropy of a set of signals is directly derived from the traditional notion. That is, the expected information quantity:

$$\begin{aligned} H(\Gamma) &= \sum_{x_1} \cdots \sum_{x_n} P(x_1, \ldots, x_n) I(O_\Gamma((x_1, \ldots, x_n)) \\ &= \sum_{\omega \in \Omega_{\mathcal{D}}} P_{d \in \mathcal{D}}(\omega = O_\Gamma(d)) I_\Gamma(d) = \frac{1}{\mathcal{D}} \sum_{d \in \mathcal{D}} I_\Gamma(d), \end{aligned}$$

which leads to Definition 3.3.

7 CONCLUSIONS

We have shown how –by starting from the Shannon-like definitions of Observational Information Quantity and Observational Entropy– we can provide a theoretically grounded explanation of phenomena that are well known results of empirical experiments. In this paper we have focused on effectiveness metrics and ranking fusion. Effectiveness can be modeled in terms of information theory –Observational Information Effectiveness (OIE), which is based on the similarity between system outputs and human assessments. OIE satisfies desirable properties that are not satisfied by traditional metrics. Moreover, our experimental results suggest that OIE captures aspects from different existing metrics. Regarding the ranking fusion problem, we have seen that, under certain assumptions, the observational information quantity outperforms single signals.

This current work has known limitations, given that the estimation of observational information quantity is not straightforward. In the near-future we plan to apply our general framework to explain other phenomena that are important in Information Retrieval, such as evaluation without relevance assessment and query performance prediction.

ACKNOWLEDGEMENTS

This research was partially supported by the Spanish Government (project Vemodalen TIN2015-71785-R) and the Australian Research Council (project LP150100242).

REFERENCES

[1] Enrique Amigó, Fernando Giner, Julio Gonzalo, and Felisa Verdejo. 2017. A Formal and Empirical Study of Unsupervised Signal Combination for Textual Similarity Tasks. In *Proceedings of ECIR'17*. 369–382.

[2] Enrique Amigó, Fernando Giner, Julio Gonzalo, and Felisa Verdejo. 2017. An Axiomatic Account of Similarity. In *Proceedings of the SIGIR'17 Workshop on Axiomatic Thinking for Information Retrieval and Related Tasks (ATIR)*.

[3] Enrique Amigó, Julio Gonzalo, and Felisa Verdejo. 2013. A General Evaluation Measure for Document Organization Tasks. In *Proceedings of SIGIR'13*. 643–652.

[4] Enrique Amigó, Damiano Spina, and Jorge Carrillo-de Albornoz. 2018. An Axiomatic Analysis of Diversity Evaluation Metrics: Introducing the Rank-Biased Utility Metric. In *Proceedings of SIGIR'18*. 625–634.

[5] Yael Anava, Anna Shtok, Oren Kurland, and Ella Rabinovich. 2016. A Probabilistic Fusion Framework. In *Proceedings of CIKM'16*. 1463–1472.

[6] Javed A. Aslam and Robert Savell. 2003. On the Effectiveness of Evaluating Retrieval Systems in the Absence of Relevance Judgments. In *Proceedings of SIGIR'03*. 361–362.

[7] David Austen-Smith and Jeffrey S. Banks. 1996. Information Aggregation, Rationality, and the Condorcet Jury Theorem. *American Political Science Review* 90, 1 (1996), 34–45.

[8] Ashraf Bah Rabiou and Ben Carterette. 2016. A Probabilistic Data Fusion Framework for Retrieval and Ranking. In *Proceedings of ICTIR'16*. 31–39.

[9] Luca Busin and Stefano Mizzaro. 2013. Axiometrics: An Axiomatic Approach to Information Retrieval Effectiveness Metrics. In *Proceedings of ICTIR'13*. 22–29.

[10] Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proceedings of CIKM'09*. 621–630.

[11] Charles L.A. Clarke, Nick Craswell, and Ian Soboroff. 2004. Overview of the TREC 2004 Terabyte Track. In *Proceedings of TREC'04*.

[12] Marquis de Condorcet. 1785. Essay on the Application of Analysis to the Probability of Majority Decisions. *Paris: Imprimerie Royale* (1785).

[13] Wagner Teixeira da Silva and Ruy Luiz Milidiú. 1993. Belief Function Model for information retrieval. *Journal of the American Society for Information Science* 44, 1 (1993), 10–18.

[14] Arthur P. Dempster. 1967. Upper and Lower Probabilities Induced by a Multivalued Mapping. *The Annals of Mathematical Statistics* (1967), 325–339.

[15] Marco Ferrante, Nicola Ferro, and Maria Maistro. 2015. Towards a Formal Framework for Utility-oriented Measurements of Retrieval Effectiveness. In *Proceedings of ICTIR'15*. New York, NY, USA, 21–30.

[16] Marco Ferrante, Nicola Ferro, and Silvia Pontarollo. 2018. A General Theory of IR Evaluation Measures. *IEEE Trans. on Knowl. and Data Eng.* (2018).

[17] Peter B. Golbus, Javed A. Aslam, and Charles L. A. Clarke. 2013. Increasing Evaluation Sensitivity to Diversity. *Inf. Retr.* 16, 4 (2013), 530–555.

[18] Lars K. Hansen and Peter Salamon. 1990. Neural Network Ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.* 12, 10 (1990), 993–1001.

[19] David A. Hull, Jan O. Pedersen, and Hinrich Schütze. 1996. Method Combination for Document Filtering. In *Proceedings of SIGIR'96*. 279–287.

[20] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20 (2002), 422–446.

[21] Serguei Kaniovski and Alexander Zaigraev. 2011. Optimal Jury Design for Homogeneous Juries with Correlated Votes. *Theory and Decision* 71, 4 (2011), 439–459.

[22] Anders Krogh and Jesper Vedelsby. 1995. Neural Network Ensembles, Cross Validation, and Active Learning. In *Proceedings of NIPS'95*. 231–238.

[23] Mounia Lalmas and Ian Ruthven. 1998. Representing and Retrieving Structured Documents using the Dempster-Shafer Theory of Evidence: Modelling and Evaluation. *Journal of Documentation* 54, 5 (1998), 529–565.

[24] Joon Ho Lee. 1995. Combining Multiple Evidence from Different Properties of Weighting Schemes. In *Proceedings of SIGIR'95*. 180–188.

[25] Gad Markovits, Anna Shtok, Oren Kurland, and David Carmel. 2012. Predicting Query Performance for Fusion-based Retrieval. In *Proceedings of CIKM'12*. 813–822.

[26] Ofer Matan. 1996. On Voting Ensembles of Classifiers. In *Proceedings of AAAI-96 Workshop on Integrating Multiple Learned Models*. 84–88.

[27] Alistair Moffat and Justin Zobel. 2008. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. Inf. Syst.* 27, 1, Article 2 (2008), 2:1–2:27 pages.

[28] Mark H. Montague and Javed A. Aslam. 2002. Condorcet Fusion for Improved Retrieval. In *Proceedings of CIKM'02*. 538–548.

[29] Rabia Nuray-Turan and Fazli Can. 2006. Automatic Ranking of Retrieval Systems using Fusion Data. 42 (05 2006), 595–614.

[30] Glenn Shafer. 1976. *A Mathematical Theory of Evidence*. Vol. 42. Princeton University Press.

[31] Claude E. Shannon. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal* 27, 3 (7 1948), 379–423.

[32] Joseph A. Shaw, Edward A. Fox, Joseph A. Shaw, and Edward A. Fox. 1994. Combination of Multiple Searches. In *Proceedings of TREC-2*. 243–252.

[33] Ian Soboroff, Charles Nicholas, and Patrick Cahan. 2001. Ranking Retrieval Systems without Relevance Judgments. In *Proceedings of SIGIR'01*. 66–73.

[34] Cornelis J. van Rijsbergen. 1979. *Information Retrieval* (2nd ed.). Butterworth-Heinemann.

[35] Christopher C. Vogt and Garrison W. Cottrell. 1998. Predicting the Performance of Linearly Combined IR Systems. In *Proceedings of SIGIR'98*. 190–196.

[36] Ellen M. Voorhees. 1999. The TREC-8 Question Answering Track Report. In *Proceedings of TREC-8*. 77–82.

A FORMAL PROOFS

Proof. **[Property 3.1]**
According to Definition 3.2, if $\forall \gamma \in \Gamma\, (\gamma(d_1) \geq \gamma(d_2))$ then:

$$\mathcal{I}_\Gamma\left(d_1\right) = -\log\left(P_{d' \in \mathcal{D}}\left(d' \geq_\Gamma d_1\right)\right) \geq$$
$$-\log\left(P_{d' \in \mathcal{D}}\left(d' \geq_\Gamma d_2\right)\right) = \mathcal{I}_\Gamma\left(d_2\right)$$

□

Proof. **[Property 3.3]**
According to Definition 3.2:

$$\mathcal{I}_{\Gamma\cup\{\gamma\}}\left(d\right) = -\log\left(P_{d' \in \mathcal{D}}\left(d' \geq_{\Gamma\cup\{\gamma\}} d\right)\right) \geq$$
$$-\log\left(P_{d' \in \mathcal{D}}\left(d' \geq_\Gamma d\right)\right) = \mathcal{I}_\Gamma\left(d\right)$$

And therefore:

$$H(\Gamma \cup \{\gamma\}) = \frac{\sum_{d\in\mathcal{D}} \mathcal{I}_{\Gamma\cup\{\gamma\}}\left(d\right)}{|\mathcal{D}|} \geq \frac{\sum_{d\in\mathcal{D}} \mathcal{I}_{\Gamma}\left(d\right)}{|\mathcal{D}|} = H\left(\Gamma\right)$$

□

Proof. **[Property 3.4]**

$$H(\{\gamma\}) = \frac{\sum_{d\in\mathcal{D}} \mathcal{I}_{\{\gamma\}}\left(d\right)}{|\mathcal{D}|} = \frac{\sum_{d\in\mathcal{D}} -\log\left(P_{d' \in \mathcal{D}}\left(d' \geq_{\{\gamma\}} d\right)\right)}{|\mathcal{D}|} =$$
$$\frac{1}{|\mathcal{D}|}\sum_{i=1}^{n} \log\left(\frac{|\mathcal{D}|}{i}\right)$$

□

Proof. **[Property 3.5]**
Being f any strict monotonic function (i.e. does not affect the ordinal relationships)

$$\mathcal{I}_{\Gamma\cup\{\gamma\}}\left(d\right) = -\log\left(P_{d' \in \mathcal{D}}\left(d' \geq_{\Gamma\cup\{\gamma\}} d\right)\right) =$$
$$-\log\left(P_{d' \in \mathcal{D}}\left(d' \geq_{\Gamma\cup\{\gamma, f(\gamma)\}} d\right)\right) = \mathcal{I}_{\Gamma\cup\{\gamma, f(\gamma)\}}\left(d\right)$$

and therefore:

$$H\left(\Gamma \cup \{\gamma\}\right) = H\left(\Gamma \cup \{\gamma, f(\gamma)\}\right)$$

□

Proof. **[Theorem 4.2 (Observational Information Evaluation Theorem)]:** Let be $\mathcal{D}$ the collection of documents. For our purposes, we can ignore the $H(g)$ from OIE given that it affects equally to every compared outputs.

Regarding the priority and deepness constraints, when swapping two contiguous documents in the ranking in concordance with the gold ($g(d_i) = 0, g(d_{i+1}) = 1$):

$$\begin{aligned}
&\mathrm{OIE}(r_{d_i\leftrightarrow d_{i+1}}) - \mathrm{OIE}(r) = H(r_{d_i\leftrightarrow d_{i+1}}) - \beta H(\{r_{d_i\leftrightarrow d_{i+1}}, g\}) - H(r) + \beta H(\{r, g\})\\
&= -\beta H(\{r_{d_i\leftrightarrow d_{i+1}}, g\}) + \beta H(\{r, g\}) \propto -H(\{r_{d_i\leftrightarrow d_{i+1}}, g\}) + H(\{r, g\})\\
&\propto \sum_{d\in\mathcal{D}} \mathcal{I}_{\{r,g\}}\left(d\right) - \sum_{d\in\mathcal{D}} \mathcal{I}_{\{r_{d_i\leftrightarrow d_{i+1}},g\}}\left(d\right)\\
&= \mathcal{I}_{\{r,g\}}\left(d_i\right) + \mathcal{I}_{\{r,g\}}\left(d_{i+1}\right) - \mathcal{I}_{\{r_{d_i\leftrightarrow d_{i+1}},g\}}\left(d_i\right) - \mathcal{I}_{\{r_{d_i\leftrightarrow d_{i+1}},g\}}\left(d_{i+1}\right)
\end{aligned}$$

Given that in both rankings the amount of relevant documents above the relevant document d_{i+1} is equal, and therefore:

$$\mathcal{I}_{\{r,g\}}\left(d_{i+1}\right) = \mathcal{I}_{\{r_{d_i\leftrightarrow d_{i+1}},g\}}\left(d_{i+1}\right)$$

Therefore, the previous expression is equivalent to:

$$\begin{aligned}
&\mathcal{I}_{\{r,g\}}\left(d_i\right) - \mathcal{I}_{\{r_{d_i\leftrightarrow d_{i+1}},g\}}\left(d_i\right)\\
&= -\log\left(P_d(d \geq_{\{g,r\}} d_i)\right) + \log(P_d(d \geq_{\{g,r_{d_i\leftrightarrow d_{i+1}}\}} d_i))\\
&= -\log\left(\frac{i}{|\mathcal{D}|}\right) + \log\left(\frac{i+1}{|\mathcal{D}|}\right) = \log\left(\frac{i+1}{i}\right)
\end{aligned}$$

Given that $\log(\frac{i+1}{i})$ is positive and monotonic regarding i, both the priority and deepness constraints are satisfied.

Regarding the threshold constraints, being D_g the set of documents annotated as relevant in the gold standard. Let r_1 a ranking which retrieves only one relevant document, its OIE is:

$$\begin{aligned}
&\mathrm{OIE}(r_1, g) \simeq H(r_1) - \beta H(\{r_1, g\})\\
&\propto \mathcal{I}_{\{r_1\}}\left(d\right) - \beta\left(\mathcal{I}_{\{r_1,g\}}\left(d\right) + \sum_{d'\in D_g-\{d\}} \mathcal{I}_{\{r_1,g\}}\left(d'\right)\right)\\
&= -\log\left(\frac{1}{|\mathcal{D}|}\right) - \beta\left(-\log\left(\frac{1}{|\mathcal{D}|}\right) - \sum_{d'\in D_g-\{d\}}\left(\log\left(\frac{|D_g|}{|\mathcal{D}|}\right)\right)\right) =\\
&= (1-\beta)\left(-\log\left(\frac{1}{|\mathcal{D}|}\right)\right) - \beta(|D_g|-1)\left(-\log\left(\frac{|D_g|}{|\mathcal{D}|}\right)\right)
\end{aligned}$$

On the other hand, let r_n a ranking which retrieves n relevant documents after n non relevant documents, then, $H(r_n)$ is $-\sum_{i=1}^{2n}\log\left(\frac{i}{|\mathcal{D}|}\right)$, and $H(\{r_n, g\})$ is proportional to:

$$\begin{aligned}
&\sum_{i=1}^{n} -\log\left(\frac{i}{|\mathcal{D}|}\right) + \sum_{i=n+1}^{2n} -\log\left(\frac{i-n}{|\mathcal{D}|}\right) + (|D_g|-n)\left(-\log\left(\frac{|D_g|}{|\mathcal{D}|}\right)\right)\\
&= -2\sum_{i=1}^{n}\log\left(\frac{i}{|\mathcal{D}|}\right) + (N_g-n)\left(-\log\left(\frac{|D_g|}{|\mathcal{D}|}\right)\right)
\end{aligned}$$

Therefore, $\mathrm{OIE}(r_n, g) = H(r_n) - \beta H(\{r_n, g\})$ can be expressed as:

$$\begin{aligned}
&-\sum_{i=1}^{2n}\log\left(\frac{i}{|\mathcal{D}|}\right) - \beta\left(-2\sum_{i=1}^{n}\log\left(\frac{i}{|\mathcal{D}|}\right) + (N_g-n)\left(-\log\left(\frac{|D_g|}{|\mathcal{D}|}\right)\right)\right)\\
&= -\sum_{i=n+1}^{2n}\log\left(\frac{i}{|\mathcal{D}|}\right) + (2\beta-1)\sum_{i=1}^{n}\log\left(\frac{i}{|\mathcal{D}|}\right) - \beta(|D_g|-n)\left(-\log\left(\frac{|D_g|}{|\mathcal{D}|}\right)\right)
\end{aligned}$$

In order to satisfy the *Deepness Threshold* constraint, the effectiveness $\mathrm{OIE}(r_n, g)$ should tend to $-\infty$ then n is extremely large. Then:

$$\begin{aligned}
&\lim_{n\to\infty} -\sum_{i=n+1}^{2n}\log\left(\frac{i}{|\mathcal{D}|}\right) + (2\beta-1)\sum_{i=1}^{n}\log\left(\frac{i}{|\mathcal{D}|}\right)\\
&\qquad - \beta(|D_g|-n)\left(-\log\left(\frac{|D_g|}{|\mathcal{D}|}\right)\right) \leq\\
&\lim_{n\to\infty} -\sum_{i=n+1}^{2n}\log\left(\frac{i}{|\mathcal{D}|}\right) + (2\beta-1)\sum_{i=1}^{n}\log\left(\frac{i}{|\mathcal{D}|}\right) <\\
&\lim_{n\to\infty} -\sum_{i=1}^{n}\log\left(\frac{i}{|\mathcal{D}|}\right) + (2\beta-1)\sum_{i=1}^{n}\log\left(\frac{i}{|\mathcal{D}|}\right) =\\
&\lim_{n\to\infty} (2\beta-2)\sum_{i=1}^{n}\log\left(\frac{i}{|\mathcal{D}|}\right) = -\infty \text{ whenever } \beta > 1
\end{aligned}$$

In order to satisfy *Closeness Deepness* constraint:

$$\begin{aligned}
&\mathrm{OIE}(r_1, g) < \mathrm{OIE}(r_n, g) \leftrightarrow (1-\beta)\left(-\log\left(\frac{1}{|\mathcal{D}|}\right)\right) - \beta(N_g-1)\left(-\log\left(\frac{|D_g|}{|\mathcal{D}|}\right)\right) <\\
&-\sum_{i=n+1}^{2n}\log\left(\frac{i}{|\mathcal{D}|}\right) + (2\beta-1)\sum_{i=1}^{n}\log\left(\frac{i}{|\mathcal{D}|}\right) - \beta(|D_g|-n)\left(-\log\left(\frac{|D_g|}{|\mathcal{D}|}\right)\right)
\end{aligned}$$

Assuming that $\beta > 1$ and $N_g > n$, we need to prove that:

$\text{OIE}(r_1, g) < \text{OIE}(r_n, g) \leftarrow$

$$(1-\beta)\left(-\log\left(\frac{N_g}{|\mathcal{D}|}\right)\right) - \beta(N_g - 1)\left(-\log\left(\frac{|D_g|}{|\mathcal{D}|}\right)\right) <$$
$$-\sum_{i=n+1}^{2n} \log\left(\frac{i}{|\mathcal{D}|}\right) + (2\beta - 1)\sum_{i=n+1}^{2n} \log\left(\frac{i}{|\mathcal{D}|}\right)$$
$$-\beta(|D_g| - n)\left(-\log\left(\frac{|D_g|}{|\mathcal{D}|}\right)\right) \Leftrightarrow$$
$$(1-\beta)\left(-\log\left(\frac{|D_g|}{|\mathcal{D}|}\right)\right) - \beta(N_g - 1)\left(-\log\left(\frac{|D_g|}{|\mathcal{D}|}\right)\right) <$$
$$-\sum_{i=n+1}^{2n} \log\left(\frac{|D_g|}{|\mathcal{D}|}\right) + (2\beta - 1)\sum_{i=n+1}^{2n} \log\left(\frac{|D_g|}{|\mathcal{D}|}\right)$$
$$-\beta(|D_g| - n)\left(-\log\left(\frac{|D_g|}{|\mathcal{D}|}\right)\right) \Leftrightarrow$$
$$(1-\beta) - \beta(|D_g| - 1) < n + (1 - 2\beta)n - \beta(|D_g| - n) \Leftrightarrow$$
$$\beta(-1 - |D_g| + 1 + 2n + |D_g| - n) < 2n - 1 \Leftrightarrow \beta < \frac{2n-1}{n}$$

Finally, regarding the *Confidence* constraint, when adding a non-relevant document d in the last position of r, the effect is that d is the only document which Observational Information Quantity changes. In addition, a non-relevant document according to g satisfies

$$\forall d' \in \mathcal{D}.\, g(d') \geq g(d)$$

which implies that $\mathcal{I}_{\{r,g\}}(d) = \mathcal{I}_{\{r\}}(d)$. Therefore, we can say that the increase of $H(\{r, g\})$ is equal than the increase of $H(\{r\})$. Therefore, according to the OIE definition, the score decreases whenever $\beta > \alpha_1$, satisfying *Confidence*.

□

Proof. **[Theorem 5.1:Ranking Mergeability]**

$$\text{OIE}\left(\mathcal{I}_{\Gamma\cup\{\Upsilon\}}\right) \geq \text{OIE}\left(\mathcal{I}_\Gamma\right) \Longleftrightarrow$$
$$H(\mathcal{I}_{\Gamma\cup\{\Upsilon\}}) - \beta H(\mathcal{I}_{\Gamma\cup\{\Upsilon\}} \cup \{g\}) \geq H(\mathcal{I}_\Gamma) - \beta H(\mathcal{I}_\Gamma \cup \{g\})$$

Assuming that $\mathcal{I}_{\Gamma\cup\{\Upsilon\}}$ and $\mathcal{I}_\Gamma$ are fine grained then:

$$H(\mathcal{I}_{\Gamma\cup\{\Upsilon\}}) = H(\mathcal{I}_\Gamma)$$

Therefore:

$$H(\mathcal{I}_{\Gamma\cup\{\Upsilon\}}) - \beta H(\mathcal{I}_{\Gamma\cup\{\Upsilon\}} \cup \{g\}) \geq H(\mathcal{I}_\Gamma) - \beta H(\mathcal{I}_\Gamma \cup \{g\}) \Longleftrightarrow$$
$$-\beta H(\mathcal{I}_{\Gamma\cup\{\Upsilon\}} \cup \{g\}) \geq -\beta H(\mathcal{I}_\Gamma \cup \{g\}) \Longleftrightarrow$$
$$-H(\mathcal{I}_{\Gamma\cup\{\Upsilon\}} \cup \{g\}) \geq -H(\mathcal{I}_\Gamma \cup \{g\}) \Longleftrightarrow$$
$$\prod_{d_0} P_d\left(d \geq_{I_{\Gamma\cup\{\Upsilon\}}} d_0, d \geq_g d_0\right) \geq \prod_{d_0} P_d\left(d \geq_{I_\Gamma} d_0, d \geq_g d_0\right) \Longleftrightarrow$$
$$\prod_{i=1..n} \frac{i}{n} \prod_{d_0} P_d\left(d \geq_{I_{\Gamma\cup\{\Upsilon\}}} d_0, d \geq_g d_0\right) \geq \prod_{i=1..n} \frac{i}{n} \prod_{d_0} P_d\left(d \geq_{I_\Gamma} d_0, d \geq_g d_0\right) \Longleftrightarrow$$
$$\prod_{d_0} \frac{P_d\left(d \geq_{I_{\Gamma\cup\{\Upsilon\}}} d_0, d \geq_g d_0\right)}{P_d\left(d \geq_{I_{\Gamma\cup\{\Upsilon\}}} d_0\right)} \geq \prod_{d_0} \frac{P_d\left(d \geq_{I_\Gamma} d_0, d \geq_g d_0\right)}{P_d\left(d \geq_{I_\Gamma} d_0\right)} \Longleftrightarrow$$
$$\prod_{d_0} P_d\left(d \geq_g d_0 \,\middle|\, d \geq_{I_{\Gamma\cup\{\Upsilon\}}} d_0\right) \geq \prod_{d_0} P_d\left(d \geq_g d_0 \,\middle|\, d \geq_{I_\Gamma} d_0\right)$$

which is true according to the Information Quantity Cumulative Evidence assumption. □