Towards a Basic Principle for Ranking Effectiveness Prediction without Human Assessments: A Preliminary Study

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

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

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

ABSTRACT

We present in this paper a preliminary study about the Observational Information Linearity (OIL) assumption as a basic principle that explains the accuracy of pseudo relevance assessments in the IR literature. The proposed model predicts the effectiveness drop curve along positions in a single ranking, the relative performance of two rankings, and converges into the traditional pseudo assessment method when considering multiple rankings and statistical independence between them.

CCS CONCEPTS

• **Information systems**;

KEYWORDS

Observational Information Theory, effectiveness predicton

ACM Reference Format:
Enrique Amigó, Stefano Mizzaro, and Damiano Spina. 2018. Towards a Basic Principle for Ranking Effectiveness Prediction without Human Assessments: A Preliminary Study. In *Proceedings of the 1st International Workshop on GeneraLization in InformAtion REtrieval (GLARE'18).* ACM, New York, NY, USA, 6 pages.

1 INTRODUCTION

Most of the research in the field of Information Retrieval (IR) is empirically based [1]. There is a general consensus this methodology is producing multiple obstacles to the community; document and query bias in data sets, human annotator and instruction bias, data set overfitting, etc. From our point of view, our community needs to complement empirical studies with other methodologies in order to advance in the development of reliable solutions for IR tasks.

Just like in others research areas, non-empirical methods require basic principles on which to base system improvements. In fact, there already exists some basic principles that have been empirically corroborated during last decades. For instance, exact word matching between query and document is a strong evidence for relevance, infrequent words or expressions have more weight as relevance predictors, documents that are clustered together behave similarly with respect to relevance to information needs (clustering hypothesis), words that share similar contexts tend to be semantically related (distributional hypothesis), etc.

In this work we focus on the prediction of document ranking effectiveness when human assessments are not available. In the literature, a common way of tackling this issue is the generation of *pseudo relevance assessments* by considering multiple system outputs. In general all these methods are based on the *Chorus* and the *Skimming* effects [14], i.e., documents retrieved by many systems in high positions are likely to be relevant. The predictability of these methods has been studied from an empirical point of view. However, their theoretical foundation is still an open issue.

Based on the *Observational Information Theory* framework [2, 3], we introduce the *Observational Information Linearity* (OIL) assumption as a basic principle that explains the effectiveness of IR system outputs. The model is able to predict effectiveness drop curve along positions in a single ranking, and the relative effectiveness of two rankings. Notice that in this basic situation, traditional pseudo assessment methods cannot be applied due to the lack of diverse rankings. In addition, the formal analysis shows that the proposed model converges into the traditional pseudo assessment methods when considering multiple rankings and statistical independence between them. In our preliminary experiments, the theoretical results are validated empirically over the GOV2 TREC data set [5].

Let us remark that this work does not attempt to provide better solutions than those that are presented in previous work; rather we aim at defining a global theoretical framework on which to base future improvements.

2 BACKGROUND

The first attempt to estimate ranking effectiveness without human assessments was proposed by Soboroff et al. [12] taking random samples from the pool of documents retrieved by the systems being evaluated. Soboroff et al. found that retaining documents returned by multiple runs for sampling improves system ranking accuracy. Wu and Crestani [15] experimented with different variants without improving substantially the results. Aslam and Savell [4] proposed comparing the top ranked document overlap peer to peer across rankings [4]. Spoerri [13] proposed avoiding the system bias by considering only one run from each research team. Sakai et al. [9] showed empirically that a simple method based on counting how many systems retrieve each document performs as well as any other existing method.

In general, the main conclusion we can infer from the literature about the generation of pseudo relevance assessments is that, given a non-biased and representative set of IR systems, the more a document is highly ranked by different systems, the more it is likely to be relevant. This is also the basis of unsupervised ranking fusion methods. In this line, Montague and Aslam [7] reported an important improvement of unsupervised combined rankings w.r.t. the most effective single ranking in multiple TREC testbeds. In

GLARE'18, 22 October, 2018, Turin, Italy
© 2018 Copyright for this paper held by its authors. Copying permitted for private and academic purposes.

addition, the need for avoiding redundant systems has been reported in the context of ranking fusion. For instance, Nuray-Turan and Can [8] reported effectiveness improvement when selecting rankings that differ from the majority voting in the ranking fusion process. Lee [6] and later Vogt and Cottrell [14] found that the best combinations were between systems that retrieve similar sets of relevant documents and dissimilar sets of non-relevant documents. Finally, Amigó et al.[2, 3] introduced the notion of *Observational Information Theory* (OIT) framework, which provides a theoretical foundation for effectiveness metrics and ranking fusion. In this paper, we use the OIT framework in order to predict the effectiveness of rankings in the absence of human assessments.

3 OBSERVATIONAL INFORMATION QUANTITY FRAMEWORK

3.1 A Basic Case

A basic principle should behave correctly under basic situations. Let us consider two rankings r_1 and r_2 in such a way that both rank documents correctly, but r_1 captures documents from the set $\mathcal{A} = \{A_1, A_2, \ldots\}$, while r_2 captures documents from both the sets $\mathcal{A}$ and $\mathcal{B} = \{B_1, B_2, \ldots\}$. Suppose that the relevant documents are $\{\underline{A_1}, \underline{A_2}, \underline{A_3}, \underline{B_1}, \underline{B_2}, \underline{B_3}\}$. Under these conditions, both rankings will have the following configuration, being r_2 more effective:

r_1 :	$\underline{A_1}$	$\underline{A_2}$	$\underline{A_3}$	A_4	A_5	A_6	A_7	A_8	…
r_2 :	$\underline{A_1}$	$\underline{B_1}$	$\underline{A_2}$	$\underline{B_2}$	$\underline{A_3}$	$\underline{B_3}$	A_4	B_4	…

Note that top documents are correctly ordered in both rankings. The difference is that r_2 is more diverse, capturing relevant documents from both $\mathcal{A}$ and $\mathcal{B}$ sets. The goal consists of predicting which is the most effective ranking without having any information about the relevance or the set to which the documents belong.

Sakai and Lin [10] proved that the simplest method of forming *pseudo-rels* based on how many systems returned each pooled document performs as well as any other existing method (see Section 2). However, in this basic case, this method would reward r_1 given that common documents (set $\mathcal{A}$) are ranked earlier.

An alternative is using voting or ranking fusion methods to estimate the relevance of documents. As an example, one of the most traditional one is the Borda Count algorithm, which assumes in this scenario that the relevance of a document is inversely correlated with the average ranking position. Documents from the set $\mathcal{B}$ are located at the bottom of r_1. Therefore, the estimated relevance for these according to Borda count will be low. In consequence, the Borda count algorithm would reward again r_1 instead of r_2. This phenomenon also happens for other voting mechanisms such as Condorcet Fuse [7] or CombMNZ [11].

Our conclusion is that when a system outperforms another system in terms of diversity, the traditional effectiveness prediction methods do not behave correctly, at least for the basic case of two rankings. We hypothesize that diversity is an aspect which distinguishes systems in terms of effectiveness. Therefore, this situation is common. Our aim is to instantiate the Observational Information Theory framework in order to capture these phenomenon.

3.2 Object Observations

The Observational Information Theoretical framework (OIT) [3] is supported by the idea of modeling the information of observations rather than objects themselves. In this sense, a document observation is interpreted as a set of signals, which can be word occurrences, meta-data, etc. In our case, a document is represented as a set of *relevance signals*, i.e., document rankings generated by IR systems. The main characteristic of OIT regarding the traditional Shannon's Information Theory is the ability to consider quantitative features. Notice that the so called *Continuous Entropy* or *Differential Entropy* models information as a continuous space of events, but objects are still represented as binary facts (statistical events that either occur or not). On the other hand, OIT models a discrete space of events (objects) associated to quantitative features. This is necessary, for instance, to represent documents as ranking positions returned by a set of IR systems.

Let $\mathcal{D}$ be a collection of documents and Γ a set of relevance signals $\{\gamma : \mathcal{D} \longrightarrow \mathcal{R}^+\}$. The core notion in OIT is the *observation*, which is an instantiation of quantitative signals.

Definition 1 (Observation). *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, $O_\Gamma(x_1, \ldots, x_n)$, under Γ is a fuzzy set of signals whose membership function corresponds with the signal values $f(\gamma_i) = x_i$, $\forall i \in \{1, \ldots, n\}$.*

As the previous definition shows, in OIT, the *observation* notion is independent from the notion of *information object*. For instance, a document is an information object, while an observation could be a set of ranking positions returned by a set of systems. However, given a set of signals, we can assume that each information object produces a certain observation.

Definition 2 (Document Observation). *An observation $O_\Gamma(d)$ of d under a set of signals Γ is an observation $O_\Gamma(\gamma_1(d), \ldots, \gamma_n(d))$.*

In summary, an observation $O_\Gamma(x_1, \ldots, x_n)$ is a fuzzy set of quantitative features (e.g., ranking positions) and each document d in the collection produces an observation $O_\Gamma(d)$ under a set of signals Γ.

3.3 Observational Information Quantity

The second core notion in OIT is measuring the information provided by observations. Just like in traditional Information Theory, the Observational Information Quantity (OIQ) measures information in terms of unlikeliness. The particularity is the use of subsumption operators between fuzzy sets.

Definition 3 (Observational Information Quantity). *The observational Information Quantity $\mathcal{I}_\Gamma(d)$ of a document d under a set of signals Γ is the minus logarithm of the probability to be subsumed by other document observations under the same signal set.*

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

According to the fuzzy set operators, OIQ can be expressed as the probability of being outperformed by other document d' for all

r_1	r_2	Doc	$\mathcal{I}_{\{r_1,r_2\}}(d)$	Improved by
$\underline{A}_1$	$\underline{A}_1$	$\underline{A}_1$	$-\log(1/N)$	$\underline{A}_1$
$\underline{A}_2$	$\underline{B}_1$	$\underline{A}_2$	$-\log(2/N)$	$\underline{A}_1$, $\underline{A}_2$
$\underline{A}_3$	$\underline{A}_2$	$\underline{A}_3$	$-\log(3/N)$	$\underline{A}_1$, $\underline{A}_2$, $\underline{B}_3$
A_4	$\underline{B}_2$	A_4	$-\log(4/N)$	...
A_5	$\underline{A}_3$	A_5	$-\log(5/N)$	...
A_6	$\underline{B}_3$	A_6	$-\log(6/N)$	...
A_7	A_4	A_7	$-\log(7/N)$	...
A_8	B_4	A_8	$-\log(8/N)$	...
...	...	$\underline{B}_1$	$-\log(2/N)$	$\underline{A}_1$, $\underline{B}_1$
		$\underline{B}_2$	$-\log(4/N)$	$\underline{A}_1$, $\underline{A}_2$, $\underline{B}_1$, $\underline{B}_2$
		$\underline{B}_3$	$-\log(6/N)$	...
		B_4	$-\log(8/N)$	...

Table 1: Exemplification for the OIT computation.

signals simultaneously:[1]

$$\mathcal{I}_\Gamma(d) = -\log\left(P_{d'\in\mathcal{D}}(\forall\gamma\in\Gamma:\gamma(d')\geq\gamma(d))\right). \quad (1)$$

Therefore, being Γ a set of rankings, and N the amount of documents in the collection, the observational information quantity of a document d can be estimated as:

$$\mathcal{I}_\Gamma(d) \simeq -\log\frac{|\{d'|\forall r\in\Gamma\,(rank(r,d')\leq rank(r,d))\}|}{N}. \quad (2)$$

In terms of ranking signals, a high OIQ implies that the document is unlikely to be outperformed by other documents for every rankings simultaneously. Table 1 exemplifies the computation of OIQ for the basic case described in the previous section. For instance, according to signals r_1 and r_2, B_2 is simultaneously outperformed by four documents: A_1, A_2, B_1 and itself. Therefore, its corresponding observational information quantity can be estimated as $\mathcal{I}_\Gamma(B_2) = \log\frac{4}{N}$, being N the total amount of documents in the collection.

3.4 Formal Properties

The three most important properties of OIQ are:

Property 1 (Signal Set Monotonicity). *OIQ is monotonic regarding the set of signals:*

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

It can be proved by considering that the more we add elements in Γ, the lower the probability of unanimous improvement in Equation 1, increasing OIQ. This means that, the more we have IR systems (rankings) the more we have information about the relevance of documents.

Property 2 (Signal Value Monotonicity). *OIQ is monotonic regarding the signal values. Whenever* $\forall\gamma\in\Gamma:\gamma(d)\geq\gamma(d')$*:*

$$\mathcal{I}_\Gamma(d)\geq\mathcal{I}_\Gamma(d').$$

That is, the higher the signals are, the less they are likely to be outperformed, which increases OIQ. In our scenario, this means that the more a document occurs in high ranking positions, the more likely to contain relevance information.

[1]Note that the inclusion operator in fuzzy sets is equivalent to say that one set outperforms another set for every feature simultaneously.

Property 3 (Non bias). *OIQ is not affected by redundant signals. Being* f *any strict monotonic function:*

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

This property is due to the fact that redundant signals do not affect unanimous outperforming. This property is the main characteristic of OIQ regarding voting mechanisms, avoiding the bias due to redundant systems. Notice that the distribution of IR approaches (rankings) depends on the preferences of system developers. However, the information quantity estimated by OIQ is not affected by rankings generated according to popular IR strategies.

4 OIQ LINEARITY ASSUMPTION

The main purpose of this paper is to provide a formal explanation for the accuracy of pseudo relevance assessments in the IR literature. On the basis of the OIT framework, we define the *Observational Information Linearity* (OIL) assumption as:

Definition 4 (Observational Information Linearity (OIL) Assumption). *Given a fixed set of IR signals (rankings), the probability of a document d to be relevant (rel(d)) is linearly correlated with its Observational Information Quantity.*

$$P(rel(d)|\Gamma)\sim\mathcal{I}_\Gamma(d).$$

Note that we are using the symbol $\sim$ to express linear correlation, that is, there exists a linear function that relates both components.

The OIL assumption can be also expressed as follows; for each set of signals Γ, there exists a constant C such that:

$$\frac{\partial P(rel(d)|\Gamma)}{\partial\mathcal{I}_\Gamma(d)}=C.$$

Under this assumption, we cannot predict the relevance of each specific document in the ranking, given that we do not know a priori this linear function or the constant C. However, we will see that it has useful implications when estimating the effectiveness of rankings in the absence of human assessments.

It is not easy to empirically show that the OIL assumption holds, as we cannot have sufficient samples of documents for each observation (i.e., the same ranking position combination across signals) to check the probability in the left part. However, we can study its direct implications. In the next subsections, we state propositions that, according to the OIL assumption, should be true when considering single, two, and more rankings, respectively. For each case, we include a small experiment using the Gov-2 collection and the topics 701 to 750 employed in the TREC 2004 Terabyte Track [5].

4.1 Implications in Single Ranking Effectiveness

The first implication determines how precision at k for a single ranking should decrease when increasing k:

Theorem 1 (OIQ Single Ranking Effectiveness). *Under the OIL assumption, given a signal r, the precision at certain ranking cutoff k is linearly correlated with the expected Observational Information Quantity under the own signal r in the first k positions:*

$$P@k(r)\sim\frac{1}{k}\sum_{i=1}^{k}\log\left(\frac{N}{i}\right).$$

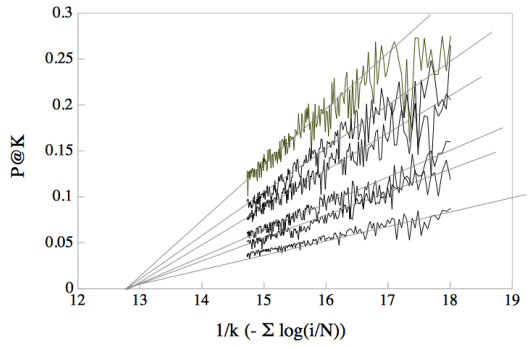

Figure 1: Checking the OIQ linearity assumption for single rankings.

Note that the linearity of the correlation is necessary to obtain the previous theorem, given that we need to aggregate probabilities across ranking positions to estimate effectiveness.

In order to check this empirically, we took the 60 official runs submitted by the participants to the track and computed their P@k for $k \in \{100, \ldots, 1000\}$ averaged across topics. Then we computed for each system the correlation between P@k and the expected OIQ score as described in Theorem 1. The reason for considering k values higher than 100 is the need for a large enough amount of samples to infer the statistics. The result was that every systems achieved a Pearson correlation between 0.8 and 0.95.

Figure 1 illustrates the result for six systems. The vertical axis represents the P@k values. The horizontal axis represenst the expected OIQ $\frac{1}{k}\sum_{i=1}^{k}\log\left(\frac{N}{i}\right)$. Each line represents a system and its P@k and expected OIQ achieved across k levels.

Although this experiment is affected by statistical noise, the figure suggests that there exists already a linear correlation, regardless of whether some systems are more effective than others. In addition, there is a point which is not directly derived from the OIL assumption: Every series converges to the same point when minimizing precision. The reason is that, when retrieving all documents from the collection (extremely high k and low expected information quantity), every ranking has a precision close to zero.

4.2 Implications when Comparing Two Rankings

Now, we study the implications of the OIL assumption when estimating the relative effectiveness of two rankings. Grounding on the OIL assumption, we can theoretically infer that the effectiveness of each ranking will be correlated with their Observational Information Quantity of documents at the top positions.

Theorem 2 (OIQ Linearity over Multiple Rankings). *Given a set of rankings, under the OIL assumption, the precision at k of each ranking $r \in R$ is linearly correlated with the expected Observational Information Quantity under the signal set R of documents above the k position in r. Formally, being R the set of rankings, and d_i^r the document at position i in the ranking r:*

$$P@k(r) \sim \frac{1}{k}\sum_{i=1}^{k}\mathcal{I}_R(d_i^r).$$

Let us consider the basic case described in Section 3.1 (see Table 1). The sum of observational information quantities at $k = 8$ for r_1 is:

$$-\log\left(\frac{1}{N}\right) - \log\left(\frac{2}{N}\right) - \ldots - \log\left(\frac{8}{N}\right)$$

which is lower than the sum achieved by r_2, that is:

$$-\log\left(\frac{1}{N}\right) - 2\cdot\log\left(\frac{2}{N}\right) - \log\left(\frac{3}{N}\right) - 2\cdot\log\left(\frac{4}{N}\right) - \log\left(\frac{6}{N}\right) - \log\left(\frac{8}{N}\right)$$

In other words, while common pseudo assessments rewards r_1 (see Section 3.1), OIQ rewards r_2 which captures document from both sets $\mathcal{A}$ and $\mathcal{B}$.

In order to check this theorem empirically, we try to predict the increase of P@k when comparing two rankings without human assessments. Using the same data set of the previous experiment and considering all possible pairs of rankings r_1 and r_2, for each ranking pair:

(1) We compute P@100 across all the topics (queries).
(2) As a first baseline, we estimate P@100 under the traditional pseudo relevance assessment (Sakai et al.'s approach [9]). That is, for each document d_i^r in the top 100 positions of r, the amount of systems (one or both) that retrieve[2] d_i^r.

$$\widehat{\text{eff}}_{\text{Pseudo}}(r) = \frac{1}{k}\sum_{i=1}^{k}\left|\left\{r \in \{r_1, r_2\} : \text{rank}(r, d_i^r) \leq 1000\right\}\right|.$$

(3) As a second baseline, we take the Borda voting method as indicator of relevance. That is, the negative average ranking position of the document across both systems.[3] We consider the average estimated relevance of documents at $k \leq 100$ as effectiveness prediction for each system.

$$\widehat{\text{rel}}_{\{r_1,r_2\}}(d) = -\,\text{avg}_{r\in\{r_1,r_2\}}\,\text{rank}(r, d)$$

$$\widehat{\text{eff}}_{\text{Borda}}(r) = \frac{1}{100}\sum_{i=1}^{100}\widehat{\text{rel}}_{\{R_1,R_2\}}\left(d_r^i\right).$$

(4) We compute their estimated effectiveness according to OIT. That is, the average OIQ score for the first 100 documents across topics. Formally, being d_r^i the i^{th} document in the ranking $r \in \{r_1, r_2\}$:

$$\widehat{\text{eff}}_{\text{OIQ}}(r) = \frac{1}{100}\sum_{i=1}^{100}\mathcal{I}_{\{r_1,r_2\}}(d_r^i),$$

where $\mathcal{I}_{\{r_1,r_2\}}(d_r^i)$ is estimated as in Equation 2 considering $N = 10,000,000$. Let us remark that in this experiment, we are considering only the two rankings in Γ.
(5) We compute the correlation across system pairs betwen the increase of P@100 ($\triangle P@100$), the increase of $\widehat{\text{eff}}_{\text{OIQ}}(r)$, $\widehat{\text{eff}}_{\text{Pseudo}}(r)$ and $\widehat{\text{eff}}_{\text{Borda}}(r)$.

[2] Here we consider that a system has retrieved the document if it occurs in the top 1000 positions.
[3] When a document is not retrieved by the system we assume a fixed position 1000.

Table 2: Correlation between P@100 improvements across ranking pairs and different prediction methods, using the two rankings as reference.

Amount of systems retrieving the document $\widehat{\text{eff}}_{\text{Pseudo}}(r)$	Average position across rankings $\widehat{\text{eff}}_{\text{Borda}}(r)$	Expected OIQ $\widehat{\text{eff}}_{\text{OIQ}}(r)$
-0.38	-0.35	0.44

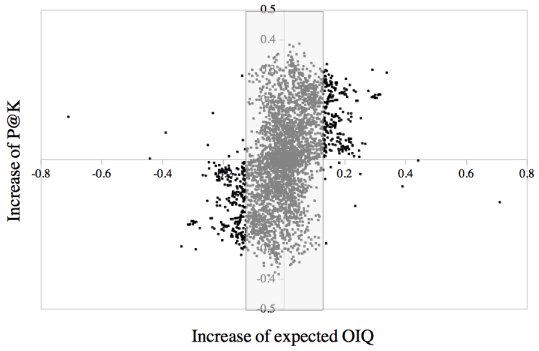

Figure 2: Checking the OIQ linearity assumption rankings pairs.

Table 3: Correlation between P@100 improvements across ranking pairs and different prediction methods using the whole set of rankings as reference.

Amount of systems retrieving the document $\widehat{\text{eff}}_{\text{Pseudo}}(r)$	Average position across rankings $\widehat{\text{eff}}_{\text{Borda}}(r)$	Expected OIQ $\widehat{\text{eff}}_{\text{OIQ}}(r)$
0.87	-0.59	0.88

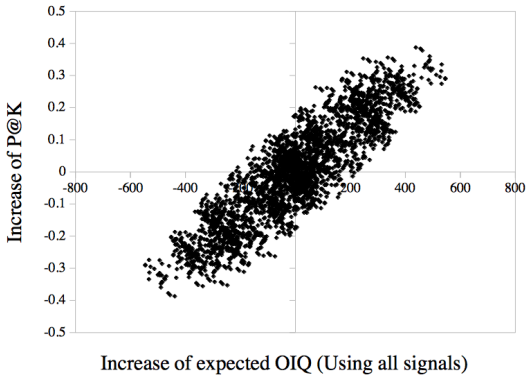

Figure 3: Checking the OIQ linearity assumption for multiple rankings rankings.

Table 2 shows the results. The expected OIQ-based estimation ($\triangle\widehat{\text{eff}}_{\text{OIQ}}(r)$) achieved a Pearson correlation with the human assessment based estimation ($\triangle P@100$) of 0.43. Similar to the basic case described in Section 3.1, the correlation achieved by the Borda-based approach was negative (-0.35), and the amount of systems retrieving the document achieves a negative correlation of -0.38.

Figure 2 illustrates the correspondence between the increase of the expected OIT and P@100. The grey rectangle shows that the P@100 increase is ensured above a certain OIT estimation difference (above 0.2). Notice that this value can change depending on the document collection size N.

4.3 Implications when Comparing Multiple Rankings

We now consider the situation in which multiple rankings are compared to each other. Unfortunately, in this case, it is not possible to estimate OIQ in a frequentist manner. Due to the combinatorial explosion, most of documents at the top of any ranking tend to be unanimously improved only by themselves. That is, there is not enough data to estimate statistics. The common solution for these situations is assuming independence. In fact, the following theorem proves that OIQ and traditional pseudo assessments (Sakai et al.'s approach [9]) tend to converge when assuming independence across rankings.

Theorem 3 (OIQ and Ranking Independence). *Being R a set of rankings, under the OIL assumption, and assuming that rankings are statistically independent, the Observational Information Quantity of a document observation is:*

$$\mathcal{I}\left(O_R(d)\right) = -\sum_{r\in R} \log(\text{rank}(r,d)/N). \tag{3}$$

In addition, being $|r_i| << N$

$$\mathcal{I}_R(d) \simeq |\{r : d \in r\}|.$$

That is, assuming that rankings are substantially shorter than the total amount of documents in the collection, there exists a convergence between the OIQ of documents and the amount of rankings returning the documents.

Experiment. In this last experiment, we repeat the previous experiment, but considering the whole set of rankings R as reference instead of the compared rankings r_1 and r_2.

Table 3 shows the results. The average position is not a good estimation of relevance, and OIT under the independence assumption converges to Sakai et al.'s approach [9], but increasing slightly its predictive power.

Figure 3 shows the correlation between the increase of expected OIT and $\triangle P@100$ in this experiment. In this case, the differences between OIQ (horizontal axis) are much larger than in the previous experiment. This is the effect of combining multiple signals in the OIQ estimation.

5 CONCLUSIONS

In this work we have presented a preliminary study about the connection between Observational Information Theory [2, 3] and the prediction of ranking relevance without human assessments. The analysis is grounded on the Observational Information Linearity

assumption, which states a correlation between the expected relevance of documents and their observational information quantity in terms of retrieval signals (rankings). This assumption allows to estimate the behavior of single rankings and the relative effectiveness of ranking pairs without having human assessments. In addition, it converges into the traditional pseudo assessment approach when assuming independence across ranking signals. Future work includes considering more data sets and more pseudo assessment approaches.

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, Hui Fang, Stefano Mizzaro, and ChengXiang Zhai. 2018. Are we on the Right Track?: An Examination of Information Retrieval Methodologies. In *Proceedings of SIGIR'18*. 997–1000.
[2] 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.
[3] Enrique Amigó, Fernando Giner, Stefano Mizzaro, and Damiano Spina. 2018. A Formal Account on Effectiveness Evaluation and Ranking Fusion. In *Proceedings of ICTIR'18*.
[4] 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.
[5] Charles L.A. Clarke, Nick Craswell, and Ian Soboroff. 2004. Overview of the TREC 2004 Terabyte Track. In *Proceedings of TREC'04*.
[6] Joon Ho Lee. 1995. Combining Multiple Evidence from Different Properties of Weighting Schemes. In *Proceedings of SIGIR'95*. 180–188.
[7] Mark H. Montague and Javed A. Aslam. 2002. Condorcet Fusion for Improved Retrieval. In *Proceedings of CIKM'02*. 538–548.
[8] Rabia Nuray-Turan and Fazli Can. 2006. Automatic Ranking of Retrieval Systems using Fusion Data. 42 (05 2006), 595–614.
[9] Tetsuya Sakai, Noriko Kando, Hideki Shima, Chuan-Jie Lin, Ruihua Song, Miho Sugimoto, and Teruko Mitamura. 2009. Ranking the NTCIR ACLIA IR4QA Systems without Relevance Assessments. *Information and Media Technologies* 4, 4 (2009), 1028–1033.
[10] Tetsuya Sakai and Chin-Yew Lin. 2010. Ranking Retrieval Systems without Relevance Assessments: Revisited. In *The Third International Workshop on Evaluating Information Access (EVIA)*. 25–33.
[11] Joseph A. Shaw and Edward A. Fox. 1994. Combination of Multiple Searches. In *Proceedings of TREC-2*. 243–252.
[12] Ian Soboroff, Charles Nicholas, and Patrick Cahan. 2001. Ranking Retrieval Systems without Relevance Judgments. In *Proceedings of SIGIR'01*. 66–73.
[13] Anselm Spoerri. 2007. Using the Structure of Overlap between Search Results to Rank Retrieval Systems without Relevance Judgments. *Information Processing and Management* 43, 4 (2007), 1059–1070.
[14] Christopher C. Vogt and Garrison W. Cottrell. 1998. Predicting the Performance of Linearly Combined IR Systems. In *Proceedings of SIGIR'98*. 190–196.
[15] Shengli Wu and Fabio Crestani. 2003. Methods for Ranking Information Retrieval Systems Without Relevance Judgments. In *Proceedings of SAC'03*. 811–816.

APPENDIX: FORMAL PROOFS

Proof. **[Theorem 1]**
The precision at k metric can be interpreted as the probability of finding a relevant document above position k. Being $rank_r(d)$ the rank position of d according to the signal r:

$$P@K(r) = \frac{|\{d \mid rel(d) \wedge rank_r(r) \leq k\}|}{k} \simeq P\big(rel(d) \,\big|\, rank_r(d) \leq k\big).$$

This can be written as:

$$\frac{1}{k}\sum_{i=1}^{k} P\big(rel(d) \,\big|\, rank_r(d) = i\big) = \sum_{i=1}^{k} \frac{P\big(rel(d) \,\big|\, rank_r(d) = i\big)}{k}. \quad (4)$$

According to the OIL assumption, this equation is linearly correlated with:

$$\sim \sum_{d:rank_r(d)\leq k} \frac{\mathcal{I}_{\{r\}}(d)}{k} = \sum_{i=1..k} \frac{\log\left(\frac{N}{i}\right)}{k}.$$

□

Proof. **[Theorem 2]**
Let us consider Equation 4 in the previous proof, the precision at K for a certain ranking can be expressed as:

$$\sum_{i=1}^{k} \frac{P\big(rel(d) \,\big|\, rank_r(d) = i\big)}{k}.$$

Then, being R a set of rankings, according to the OIL assumption, this is linearly correlated with:

$$\sim \sum_{d:rank_r(d)\leq k} \frac{\mathcal{I}_R(d)}{k} = \sum_{i=1}^{k} \frac{\mathcal{I}_R(d_i^r)}{k}.$$

□

Proof. **[Theorem 3]**
For this proof, we will consider each ranking r as a subset of sorted documents from $\mathcal{D}$, in such a way that $d \in r$ represents that the document is retrieved in r and $|r|$ corresponds with the ranking size. We assume that documents out of the ranking achieve a fixed extremely low $r(d)$.

Being R a set of ranking signals, according to Equation 1, OIQ can be expressed as:

$$\mathcal{I}_R(d) = -\log\left(P_{d'\in\mathcal{D}}\left(\forall r \in R : r(d') \geq r(d)\right)\right).$$

Assuming statistical independence across rankings, this is equivalent to:

$$-\sum_{r\in R} \log\left(P_{d'\in\mathcal{D}}\left(r(d') \geq r(d)\right)\right) \simeq -\sum_{r\in R} \log(rank(r,d)/N),$$

which corresponds with the first part of the theorem. Now, discarding ranking in which the document does not appear, for which

$$log\left(P_{d'\in\mathcal{D}}\left(r(d') \geq r(d)\right)\right) = 0,$$

then the previous equation is equivalent to:

$$-\sum_{r\in R:d\in r} \log(rank(r,d)/N).$$

Assuming $N >> |r|$ for all r in R, then $-log(rank(r,d)/N)$ is extremely high for document in the ranking. Therefore,

$$\sum_{r\in R:d\in r} -\log(rank(r,d)/N) \simeq |\{r \mid d \in r\}|$$

as the theorem states.

□