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

Enrique Amigó 1  Stefano Mizzaro 2  Damiano Spina 3

1UNED NLP & IR Group
2University of Udine
3RMIT University

October 22, 2018
Overview

1. Background
   - The Problem
   - Statements and Observed Phenomena
   - Partial Explanations for the Relevance Predictability

2. Proposal
   - Observational Information Quantity (OIQ)
   - Obs. Inf. Linearity Assumption (OIL)

3. Implications
   - Relevance decay across ranking positions.
   - Ranking Effectiveness given a Ranking Set
   - Comparing ranking pairs.

4. Conclusions

shortname  A Basic Principle for Effectiveness Predictability
Background

The Problem
Statements and Observed Phenomena
Partial Explanations for the Relevance Predictability

A Basic Principle for Effectiveness Predictability
Ranking Fusion vs. Effectiveness Prediction without Human Assessments (Pseudo-QRels).

Two faces of the same coin [Nuray and Can, 2006].

A Basic Principle for Effectiveness Predictability
Ranking Fusion vs. Effectiveness Prediction without Human Assessments (Pseudo-QRels).

Two faces of the same coin [Nuray and Can, 2006].

The goal: Predicting the relevance $\text{Rel}(d)$ of documents given a set $\mathcal{R}$ of rankings.

\[ \text{Ranking Fusion} \quad \Rightarrow \quad \text{Pseudo-qrels} \]

\[ r_1 \quad r_2 \quad r_3 \]

\[ \rightarrow \quad \text{High Rel}(d) \]

\[ \Rightarrow \quad \text{Rel}(d) \]
Probability Ranking Principle: “Retrieval systems return documents ranked in order of decreasing probability of relevance to the user” [Rijsbergen, 1979].


- **Probability Ranking Principle**: “Retrieval systems return documents ranked in order of decreasing probability of relevance to the user” [Rijsbergen, 1979].

- **Effectiveness Convergence**: “Different approaches return non overlapped document sets, but they tend to be similarly effective.”

---

Predictive Power of Unsupervised Approaches:

- **Pseudo-qrels:** Human annotations can be replaced by docs. retrieved by systems [Soborof et al 2001].
Statements and Observed Phenomena (II)

- **Predictive Power of Unsupervised Approaches:**
  - **Pseudo-qrels:** Human annotations can be replaced by docs. retrieved by systems [Soborof et al 2001].

- **Simple approaches achieve competitive results:**
  - **Ranking fusion:** CombMNZ=“Sum of scores” × “How many systems retrieve it” [Lee, 1997].
  - **Pseudo-qrels:** nruns=”How many systems retrieve it” [Sakai and Lin 2010].
The Role of Diversity:

- **Ranking fusion:** “Effectiveness increases when relevant documents are ranked in a different fashion” [Vogt and Garrison, 1998].

- **Pseudo-qrels:** Selecting systems that differ from the norm [Nuray and Can, 2006] or selecting one system from each team [Spoerri 2008] improves prediction.
Common Points

- Three main variables:
  - How many systems retrieve it?
Common Points

- **Three main variables:**
  - How many systems retrieve it?
  - In what ranking **positions** it is retrieved?
Common Points

- **Three main variables:**
  - How *many* systems retrieve it?
  - In what ranking *positions* it is retrieved?
  - How *diverse* are the systems that retrieve it?
Partial Explanations for the Relevance Predictability

**Skimming Effect**: Dissimilar relevant docs. at the top.

[Belkin et al., 1993, Vogt and Garrison, 1998]
Partial Explanations for the Relevance Predictability

- **Skimming Effect**: Dissimilar relevant docs. at the top.
  
  [Belkin et al., 1993, Vogt and Garrison, 1998]

- **Chorus Effect**: Similar relevant docs. at the top but dissimilar irrelevant docs.
  
Partial Explanations for the Relevance Predictability

- **Skimming Effect**: Dissimilar relevant docs. at the top.
  
  [Belkin et al., 1993, Vogt and Garrison, 1998]

- **Chorus Effect**: Similar relevant docs. at the top but dissimilar irrelevant docs.
  

- **Compatibility Constraint**: Linear correlation between relevance and scores. [Gerani et al. 2012]
Partial Explanations for the Relevance Predictability

- **Skimming Effect**: Dissimilar relevant docs. at the top.
  [Belkin et al., 1993, Vogt and Garrison, 1998]

- **Chorus Effect**: Similar relevant docs. at the top but dissimilar irrelevant docs.

- **Compatibility Constraint**: Linear correlation between relevance and scores. [Gerani et al. 2012]

There is yet no satisfactory theoretical explanation for these phenomena.
Proposal: the Observational Information Quantity and the Observational Linearity Assumption
Given by the probability of being outscored by other documents in all rankings.

\[ \mathcal{I}_\mathcal{R}(d) = -\log \left( P_{d' \in \mathcal{D}} \left( \forall r \in \mathcal{R}. \text{Rank}_r(d') \leq \text{Rank}_r(d) \right) \right). \]
Assuming that rankings are statistically independent:

\[ I_R(d) = -\sum_{r \in R} I_r(d) \simeq -\sum_{r \in R} \log \left( \frac{\text{Rank}_r(d)}{N} \right). \]
Formal Properties vs. Predictive Variables

- **How many** systems retrieve it?

\[ \mathcal{I}_\mathcal{R}(d) \text{ is monotonic regarding the set } \mathcal{R}. \]

\[ \mathcal{I}_{\mathcal{R} \cup \{r\}}(d) \geq \mathcal{I}_\mathcal{R}(d) \]
Formal Properties vs. Predictive Variables

- **How many** systems retrieve it?

  \[ I_{\mathcal{R}}(d) \] is monotonic regarding the set \( \mathcal{R} \).

  \[ I_{\mathcal{R} \cup \{r\}}(d) \geq I_{\mathcal{R}}(d) \]

- **In what ranking positions** it is retrieved?

  \[ I_{\mathcal{R}}(d) \] is monotonic regarding \( \text{Rank}_r(d) \).

  \[ \forall r \in \mathcal{R}. \text{Rank}_r(d_1) \leq \text{Rank}_r(d_2) \implies I_{\mathcal{R}}(d_1) \geq I_{\mathcal{R}}(d_2). \]
Formal Properties vs. Predictive Variables

- **How many** systems retrieve it?

  \[ I_R(d) \text{ is monotonic regarding the set } R. \]

  \[ I_{R \cup \{r\}}(d) \geq I_R(d) \]

- **In what ranking positions** it is retrieved?

  \[ I_R(d) \text{ is monotonic regarding } \text{Rank}_r(d). \]

  \[ \forall r \in R. \text{Rank}_r(d_1) \leq \text{Rank}_r(d_2) \implies I_R(d_1) \geq I_R(d_2). \]

- **How diverse** are the systems that retrieve it?

  \[ I_R(d) \text{ is not affected by redundant rankings in } R. \]

  \[ I_{R \cup \{r\}}(d) = I_{R \cup \{r,r\}}(d) \]
Observational Information Linearity Assumption (OIL)
The probability of relevance and OIQ are **linearly** correlated.

\[
\frac{\partial P(\text{rel}(d) \, | \, \mathcal{R})}{\partial I_{\mathcal{R}}(d)} = C
\]
The probability of relevance and OIQ are **linearly** correlated.

\[
\frac{\partial P\left(\text{rel}(d) \mid \mathcal{R}\right)}{\partial I_{\mathcal{R}}(d)} = C
\]

It is equivalent to say that \( I_{\mathcal{R}}(d) \) satisfies the Compatibility Constraint [Gerani et al. 2012].
A Basic Principle for Effectiveness Predictability

\[
\frac{\partial P(\text{rel}(d)| \mathcal{R})}{\partial I_{\mathcal{R}}(d)} = C
\]
OIL Implications

- Predicting document relevance decay:

\[ \{r\} \rightarrow \frac{\partial \text{rel}(d^r_i)}{\partial i} \]
OIL Implications

- Predicting document relevance decay:

\[ \{ r \} \rightarrow \frac{\partial rel(d^r_i)}{\partial i} \]

- Predicting ranking effectiveness (given a set of rankings):

\[ \{ r_1, r_2, \ldots, r_n \} \rightarrow Eff(r_i) \]
OIL Implications

- Predicting document relevance decay:
  \[ \{r\} \rightarrow \frac{\partial \text{rel}(d^r_i)}{\partial i} \]

- Predicting ranking effectiveness (given a set of rankings):
  \[ \{r_1, r_2, \ldots, r_n\} \rightarrow \text{Eff}(r_i) \]

- Comparing two rankings without any other reference.
  \[ \{r_1, r_2\} \rightarrow \text{Eff}(r_1) > \text{Eff}(r_2) \]
Predicting document relevance decay:

\[
\{r\} \rightarrow \frac{\partial \text{rel}(d'_i)}{\partial i}
\]
Under a single ranking OIQ depends on the rank position:

\[ I_r(d_i^r) = -\log \left( \frac{i}{N} \right) \]
Under a single ranking OIQ depends on the rank position:

\[ I_r(d_i^r) = -\log \left( \frac{i}{N} \right) \]

Therefore, if the OIL assumption is true:

**Theorem (OIQ Single Ranking Effectiveness)**

*Precision at k is linearly correlated with a function of k:*

\[ \frac{\partial P@k(r)}{\partial f(k)} = C \text{ where } f(k) = \frac{1}{k} \sum_{i=1}^{k} \log \left( \frac{N}{i} \right). \]
Under a single ranking OIQ depends on the rank position:

\[ I_r(d^r_i) = -\log \left( \frac{i}{N} \right) \]

Therefore, if the OIL assumption is true:

**Theorem (OIQ Single Ranking Effectiveness)**

*Precision at k is linearly correlated with a function of k:*

\[ \frac{\partial P@k(r)}{\partial f(k)} = C \text{ where } f(k) = \frac{1}{k} \sum_{i=1}^{k} \log \left( \frac{N}{i} \right). \]

*A stricter version of the Probability Ranking Principle.*
Predicting document relevance decay

Figure: Implication of OIL for single rankings in TREC 2004.
Predicting document relevance decay

Figure: Implication of OIL for single rankings in four IR data sets.
Conclusion 1: OIL is satisfied by systems for a single ranking, and underlies the Probability Ranking Principle.
Predicting ranking effectiveness:

\[ \{r_1, r_2, \ldots, r_n\} \rightarrow \text{Eff}(r_i) \]
Theorem (OIL Based effectiveness Prediction)

Effectiveness of rankings at $k$ are linearly correlated with the aggregation of their OIQ’s above $k$ position.

\[
\frac{\partial P@k(r)}{\partial f(r)} = C \quad \text{where} \quad f(r) = \sum_{i=1}^{k} I_R(d_i^r).
\]
Ranking Effectiveness given Multiple Rankings

\[ P@k(r) \sim \sum_{i=1}^{k} I_R(d_i^r). \]
Ranking Effectiveness given Multiple Rankings

\[ P@k(r) \sim \sum_{i=1}^{k} I_R(d_i^r). \]
Ranking Effectiveness given Multiple Rankings

\[ P@k(r) \sim \sum_{i=1}^{k} I_R(d_i^r). \]
Ranking Effectiveness given Multiple Rankings

\[ P@k(r) \sim \sum_{i=1}^{k} I_R(d_i^r). \]

A Basic Principle for Effectiveness Predictability
A Basic Principle for Effectiveness Predictability

Relevance decay across ranking positions.

Ranking Effectiveness given a Ranking Set

Comparing ranking pairs.

Ranking Effectiveness given Multiple Rankings

\[ P@k(r) \sim \sum_{i=1}^{k} \mathcal{I}_R(d_r^i). \]

\[ P@3(r_1) = P@3(r_2) > P@3(r_3) \]
Experiment: Predicting ranking effectiveness.

Baselines:
- **nruns**: $Rel(d) \simeq \text{Amount of systems that retrieve it.}$
- **Sakai**: $Rel(d) \simeq \text{Amount of systems} + \frac{1}{1+\text{Average rank}}$
- **Nuray**: $Rel(d) \simeq \text{wins} + \frac{1}{1+\text{loses}}$
- **OIQ**: $Rel(d) \simeq \mathcal{I}_R(d)$
- **OIQ (independence)**: $Rel(d) \simeq \sum_{r \in \mathcal{R}} \mathcal{I}_r(d)$
Experiment: Predicting ranking effectiveness.

Baselines:
- **nruns**: \( \text{Rel}(d) \approx \) Amount of systems that retrieve it.
- **Sakai**: \( \text{Rel}(d) \approx \) Amount of systems + \( \frac{1}{1 + \text{Average rank}} \)
- **Nuray**: \( \text{Rel}(d) \approx \) wins + \( \frac{1}{1 + \text{loses}} \)
- **OIQ**: \( \text{Rel}(d) \approx \mathcal{I}_R(d) \)
- **OIQ (independence)**: \( \text{Rel}(d) \approx \sum_{r \in \mathcal{R}} \mathcal{I}_r(d) \)

**nruns approaches OIQ\text{\_ind}** under a wide set of systems and a huge document collection.

\[
N \gg |r| \land |\mathcal{R}| \gg 1 \implies \sum_{r \in \mathcal{R}} \mathcal{I}_r(d) \sim |\{ r : d \in r \}|.
\]
Results.

Pearson correlation between predicted and “real” P@k.

Baseline approaches behave similarly, as well as OIQ_{ind}, while the original OIQ is not suitable for many rankings (Is that due to statistic limitations?)
Results.

Pearson correlation between predicted and “real” DCG.

Baseline approaches behave similarly, as well as OIQ\textsubscript{ind}, while the original OIQ is not suitable for many rankings (Is that due to statistic limitations?)
An extreme case: Comparing two rankings without any other reference.

\[ \{r_1, r_2\} \rightarrow Eff(r_1) > Eff(r_2) \]
Predicting P@100 for system pairs.

Correlation between differences according to the estimated vs. real effectiveness.

Conclusion 3: Under a few rankings, OIQ is more predictive than baseline approaches, excepting for TREC 8.
Predicting DCG for system pairs.

Correlation between differences according to the estimated vs. real effectiveness.

Conclusion 3: Under a few rankings, OIQ is more predictive than baseline approaches, excepting for TREC 8.
Conclusions. (Theoretical)

- OIQ captures the three main predictive variables reported in the literature: system amount, ranking position and diversity.
OIQ captures the three main predictive variables reported in the literature: system amount, ranking position and diversity.

The Observational Information Linearity assumption (OIL):
- Explains the Probabilistic Ranking Principle (Rijsbergen)
Conclusions. (Theoretical)

- OIQ captures the three main predictive variables reported in the literature: system amount, ranking position and diversity.
- The Observational Information Linearity assumption (OIL):
  - Explains the Probabilistic Ranking Principle (Rijsbergen)
  - Has a correspondence with the Compatibility Constraint (Gerani et al)
OIQ captures the three main predictive variables reported in the literature: system amount, ranking position and diversity.

The Observational Information Linearity assumption (OIL):

- Explains the Probabilistic Ranking Principle (Rijsbergen)
- Has a correspondence with the Compatibility Constraint (Gerani et al)
- Corresponds with nruns when assuming independent rankings (Strong baseline according to Sakai)
Conclusions. (Empirical)

- OIL is corroborated for single rankings in 5 data sets.
Conclusions. (Empirical)

- OIL is corroborated for single rankings in 5 data sets.
- Most of baseline approaches are similarly predictive than OIQ when considering multiple systems and assuming independence.
Conclusions. (Empirical)

- OIL is corroborated for single rankings in 5 data sets.
- Most of baseline approaches are similarly predictive than OIQ when considering multiple systems and assuming independence.
- When considering two rankings, OIQ outperforms baselines excepting for TREC8.
We hypothesize that the Observational Information Linearity assumption explains the Probability Ranking Principle and the predictive power of relevance estimators. The Challenge is estimating the probability of unanimous outscoring in OIQ.
Future work

- First pending task: What happens with TREC8? The effect of pooling in the campaign?
Future work

- First pending task: What happens with TREC8? The effect of pooling in the campaign?
- Solving OIQ estimation: Using techniques for reducing the amount of rankings.
Future work

- First pending task: What happens with TREC8? The effect of pooling in the campaign?
- Solving OIQ estimation: Using techniques for reducing the amount of rankings.
Thanks for your attention
References


References

Luca Busin and Stefano Mizzaro (2013)

Combining Evidence from Multiple Searches In: The First Text REtrieval Conference (TREC-1), Gaitherburg, MD, USA, pp. 319328

Gerani, S., Zhai, C. and Crestani, F. (2012)
Score Transformation in Linear Combination for Multi-criteria Relevance. Advances in Information Retrieval.

Joon Ho Lee (1995)

Joon Ho Lee (1997)
References

Mark H. Montague and Javed A. Aslam (2002)

Ng, K. B. (1998)


References

- Van Rijsbergen, Cornelis J. (1997)  
  Information Retrieval

  Ranking Retrieval Systems without Relevance Assessments- Revisited.  
  *EVIA 2010*

- Saracevic, T., Kantor, P. (1988)  
  A study of information seeking and retrieving, iii: Searchers, searches, overlap.  
  *Journal of the American Society of Information Science* 39(3), 197-216
References

Soboroff, I. and Nicholas, C. and Cahan, P. (2001)
Ranking retrieval systems without relevance judgments. *SIGIR’01*

Authority and Ranking Effects in Data Fusion. *Journal of the American Society for Information Science and Technology* 3(59), 450-460


Predicting the Performance of Linearly Combined IR Systems. *In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*

Wu, Shengli (2012)
Data Fusion in Information Retrieval.