

Intrinsic Dimensionality in IR

陳宣穎

Paper:

Vincent Claveau, Indiscriminateness in representation spaces of terms and documents, ECIR 2018

Local Intrinsic Dimensionality

- ❖ to portray the neighboring documents of a query within distance
- ❖ how many variables are needed to generate a good approximation of the query
- ❖ as an interpretation of indiscriminateness of dataset

Local Intrinsic Dimensionality in IR

- ❖ Assume that there are two balls with center c_1 and c_2 and radius of ϵ_1 and ϵ_2 :

The ratio between the volumes of these balls can be expressed as:

$$\frac{\text{volume}(B(x, \epsilon_1))}{\text{volume}(B(x, \epsilon_2))} = \left(\frac{\epsilon_1}{\epsilon_2}\right)^m$$

$$m = \frac{\ln(\text{volume}(B(x, \epsilon_1))) - \ln(\text{volume}(B(x, \epsilon_2)))}{\ln \epsilon_1 - \ln \epsilon_2}$$

Local Intrinsic Dimensionality in IR

- ❖ Replace the volume itself with the number of points

$$\hat{m} = \frac{\ln |B(x, \epsilon_1)| - \ln |B(x, \epsilon_2)|}{\ln \epsilon_1 - \ln \epsilon_2}$$

- ❖ Apply the RSV(Retrieval Status Value) to the function, and

$$RSV(q, d) = \sum_{t \in q} w_q(t) \cdot w_d(t)$$

Distribution of the Documents

- ❖ Documents fall in the space with two thresholds values ϵ_1 and ϵ_2 ($\epsilon_1 \geq \epsilon_2$).

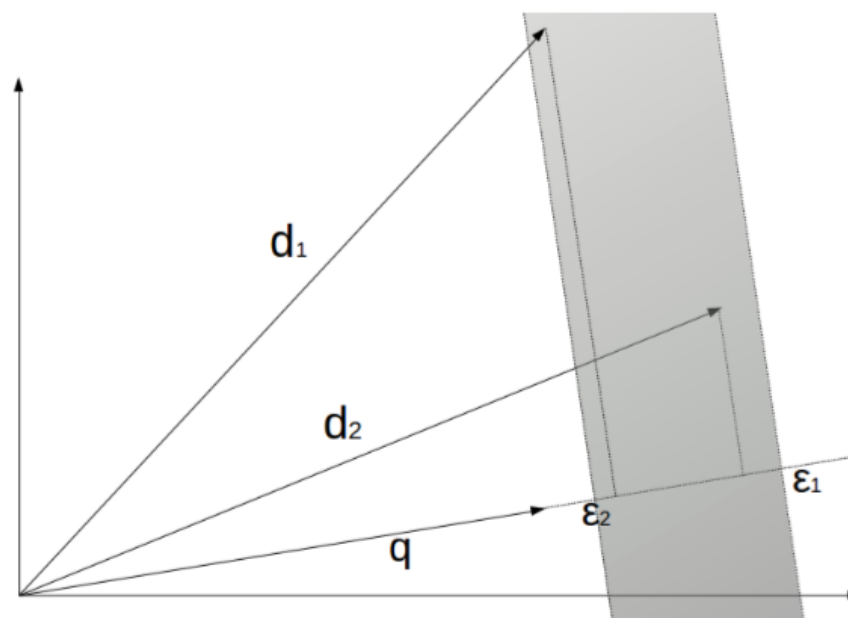


Fig. 3: In gray: portion of space defined by the set of points whose scalar products with a normed vector q lie between ϵ_1 and ϵ_2

Distribution of the Documents

- ❖ The close documents may have any distance because of no normalization.

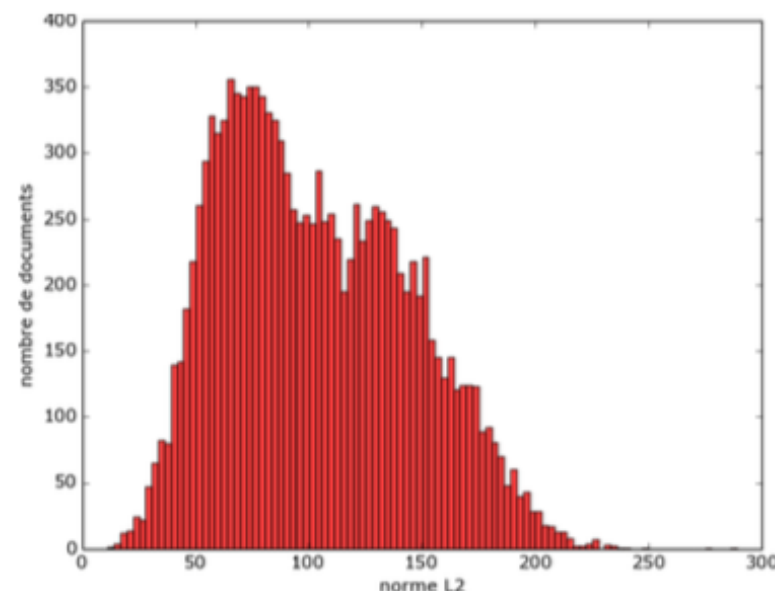


Fig. 1: Distribution of the L2 norms of documents in Tipster collection under BM25+ (modified version of BM25 proposed by [16])

Estimate with Power Law

- ❖ By repartition the distance between q and documents, we can estimate the intrinsic dimension.
- ❖ The distribution of documents follow the Power Law, so we can interpret intrinsic dimension α as the exponent.
- ❖ x represents the RSV score

$$f(x) = \lambda x^{-\alpha} \text{ with } \lambda \text{ a constant and } \alpha > 1$$

Estimate with Power Law

- ❖ We can set a threshold to acquire α with the top n RSV scores

$$\hat{\alpha} = 1 + n \cdot \left(\sum_{i=1}^n \ln \frac{x_i}{x_{min}} \right)^{-1}$$

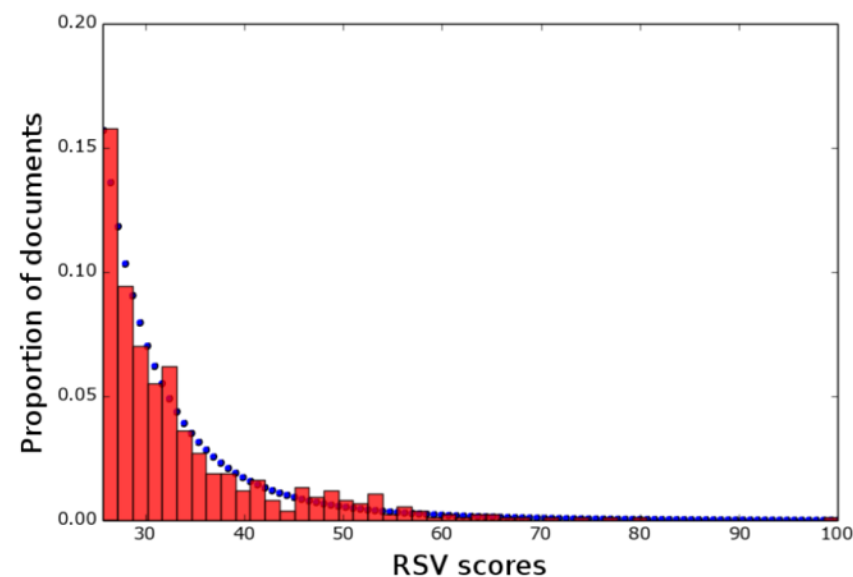


Fig. 4: Example RSV values repartition (red histogram) and the corresponding Power Law (blue) obtained with log-likelihood estimate of α from the RSV values

Experiment in IR

- ❖ The distribution of documents determines the retrieval precision, so we can assess the difficulty of query by exploring the correlation between the α and the AP(Average Precision).
- ❖ Low retrieval happens with high indiscriminateness around query.

coefficient	Value	p-value
Pearson r	-0.7150	$5.43e^{-09}$
Spearman ρ	-0.7753	$3.82e^{-11}$
Kendall τ	-0.5755	$3.69e^{-09}$

Table 2: Correlations (and their associated p-values) between AP and index α on Tipster with a BM25+ model

coefficient	Value	p-value
Pearson r	-0.4919	$9.85e^{-08}$
Spearman ρ	-0.6141	$3.26e^{-12}$
Kendall τ	-0.4494	$1.14e^{-11}$

Table 3: Correlations (and their associated p-values) between AP and index α on OHSUMED with a Dirichlet LM ($\mu = 1000$)

Experiment in Query Expansion

- ❖ Adding the closest semantic term to the original query
- ❖ We can estimate the α to obtain the properties of the word space, then use it to filter the expansions.
- ❖ The author set two experiments with different filters for the expansion.

Experiment in Query Expansion

- ❖ Two filters:

- ❖ Filter 1

- ❖ compute the α for each word of the query and pick out the word with α lower than the threshold (average α of the query in this case)

- ❖ Filter 2

- ❖ filter the word by filter 1 first, then choose the words with α below a certain for a second time

Experiment in Query Expansion

- ❖ With close examination, words with high α are polysemic or common, such as use, way, young, etc.

	MAP	R-Prec	P@5	P@10	P@50	P@100
No expansion	21.78	30.93	92.80	89.40	79.60	70.48
with expansion	+13.80	+9.58	+2.16	+4.03	+5.58	+8.26
with expansion + Filter 1	+16.22	+10.78	+3.02	+4.47	+9.20	+12.51
with expansion + Filter 1 & 2	+22.83	+13.00	+2.56	+6.31	+14.10	+21.39

Table 4: Relative performance gain (%) on Tipster with query expansion with and without filtering; spectral lexicon

	MAP	R-Prec	P@5	P@10	P@50	P@100
No expansion	21.78	30.93	92.80	89.40	79.60	70.48
with expansion	+13.52	+9.50	+2.59	+3.36	+8.29	+9.99
with expansion + Filter 1	+15.73	+9.27	+2.22	+4.96	+9.63	+14.41
with expansion + Filter 2	+20.76	+13.63	+3.88	+5.82	+10.15	+14.27

Table 5: Relative performance gain (%) on Tipster with query expansion with and without filtering; Word2Vec

Conclusion

- ❖ Leveraging the notion of intrinsic dimensionality in place of the distance to evaluate the similarity in IR raises question.
- ❖ Practically, this technique can help form a better query for online search by suggesting more precise words when a word is typed in and its α is rather high.

Thank you for your attention