From Information Theory to Term Weighting

Alternatives to Classic BM25-IDF based on a New Information Theoretical Framework

Weimao Ke

wk@drexel.edu

College of Computing and Informatics

Drexel University, Philadelphia, U.S.A.

December 18, 2022 @ IEEE Big Data Conference

Outline

- Background: TF*IDF and BM25
- Theory: Discounted Least Information Theory of Entropy
- Application: DLITE for Term Weighting
- Experimental Setup
- Results and Finding
- Conclusion

This research was partially supported by IMLS grant #RE-246450-OLS-20 (LEADING) and NSF grant #1741306.

Background: TF*IDF

Classic TF*IDF:

- Term Frequency (tf_{dt}): # occurrences of term t in document d
- Document Frequency (n_t): # docs containing term t

```
Okapi BM25 is a version of TF*IDF and
the default scoring in Elastic Search.
```

Background: TF

Variants of TF weight, $w_{dt}^{TF} =$:

1. Raw frequency: tf_{dt}
2. Normalization with logarithm: $\log tf_{dt} + 1$
3. Normalization with saturation: $\frac{tf_{dt}}{k+tf_{dt}}$
4. Document length normalization: $\frac{tf_{dt}}{l_d}$

`In BM25, the TF component is a combination of #3 (saturation) and #4 (doc length normalization).`

Background: IDF

Classic IDF formula: $w_t^{IDF} = \log \frac{N}{n_t}$, where N is the total # of docs.

Three perspectives on IDF (Inverse Document Frequency):

1. Heuristic (Salton): The **inverse** relation between a term's informativeness and how commonly (or rarely) it appears.
2. Probabilistic (Robertson & Sparck Jones): The estimate of a term's contribution to the **log-likelihood** (odds) of document relevance, i.e. retrieval status value (RSV) in IR ranking.
3. Information theoretical (Aizawa): The amount of information in a term measured by **KL Divergence** or conditional entropy.

Salton

Robertson

Sparck Jones

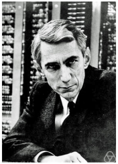
Aizawa

Background: Entropy and KL Divergence

Kullback-Leibler (KL) Divergence, a.k.a. relative entropy:

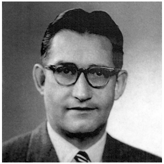
$$KL(P||Q) = \sum_{x \in X} p_x \log \frac{p_x}{q_x}$$

measures the amount of discrimination information (in Shannon Entropy) for distributions P and Q.



Shannon

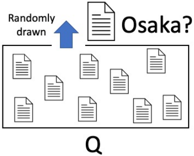
Kullback



Leibler

Background: IDF as KL Divergence

P and Q in a text collection:



- $q_t = n_t/N$ be the probability of observing term t in a randomly drawn document.
- $q'_t = 1 - q_t$ the probability of NOT observing the term.

- $p_t = 1$ be in the probability of observing term t in a document containing the term.
- $p'_t = 0$ the probability of NOT observing it in that document.

It can be shown that:

$$KL(P_t||Q_t) = \log \frac{N}{n_t}$$

`IDF captures the discriminative power (information) of a term based on KL Divergence.`

Background: Issues with KL-based IDF

Theoretical properties of KL divergence:

1. Not a metric distance
   - Not symmetric, $KL(P||Q) \neq KL(Q||P)$
   - Not satisfying triangular inequality
   - Hard to interpret how scores add up (sum)
2. Unbounded, can have infinite values

Implications on IDF w_t^{IDF}:

1. Interpretation of $\sum_t w_t^{IDF}$?
2. Can have relatively large w_t^{IDF} for a rare term t.

Example: "The Mochi is **sooooooooooooooo** tasty."

Theory: LIT

Least Information Theory, $LIT(P, Q)$:

$$= \sum_{x \in X} \int_{p_x}^{q_x} -\log p \; dp$$

$$= \sum_{x \in X} \left| p_x(1 - \ln p_x) - q_x(1 - \ln q_x) \right|$$

- Metric distance, bounded
- Great results in IR, clustering, classification, and many others.
- Missing an important information-theoretic property
  - NOT satisfying the breakdown rule, i.e. the LIT of an ensemble $\neq$ the weighted sum of LITs in the sub-systems.
  - i.e. $lit(x \cdot p, x \cdot q) \neq x \cdot lit(p, q)$

Ke (2012, 2013), Gong & Ke (2013)
Gong (2015), Ke (2015, 2017), Du & Ke (2018)

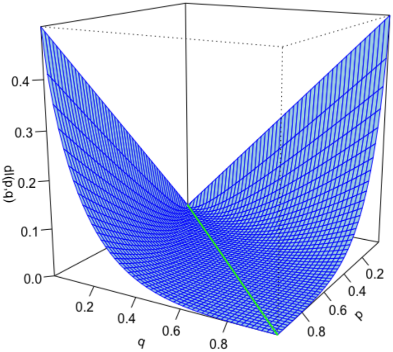
Theory: DLITE

We introduce an entropy discount $\Delta_H(P, Q)$:

$$= \sum_{x \in X} \left| p_x - q_x \right| \frac{\int_{p_x}^{q_x} -p \log p \ dp}{\int_{p_x}^{q_x} x \ dx}$$

Discounted LIT of Entropy, DLITE (pronounced *delight*) is:

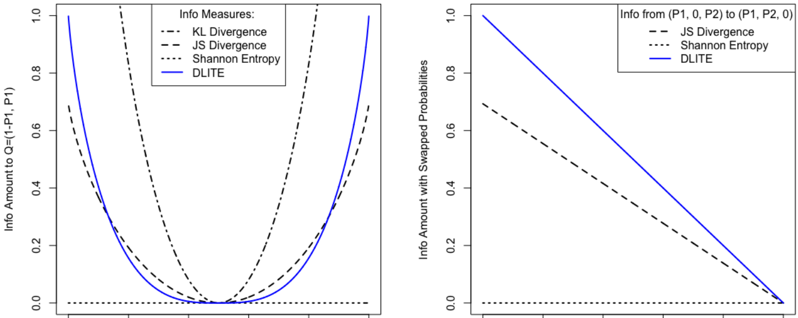
$$DL(P, Q) = LIT(P, Q) - \Delta_H(P, Q)$$



Theory: DLITE Properties (Highlights)

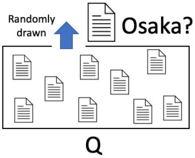
With the entropy discount, DLITE:

- Satisfies several information-theoretic properties, including:
  - The breakdown rule: The DLITE of an ensemble = the weighted sum of DLITEs in the sub-systems.
  - i.e. $dl(x \cdot p, x \cdot q) = x \cdot dl(p, q)$
- Non-negative, bounded in $[0, 1]$, symmetric
- $DL(P, Q) = 0$ only when P and Q are identical.
- $\sqrt[3]{DLITE}$ satisfies triangular inequality (metric distance).



Application: DLITE Alternative to IDF

Remember P_t and Q_t in a text collection:



- $q_t = n_t / N$ be the probability of observing term t in a randomly drawn document.
- $q'_t = 1 - q_t$ the probability of NOT observing the term.

- $p_t = 1$ be in the probability of observing term t in a document containing the term.
- $p'_t = 0$ the probability of NOT observing it in that document.

DLITE's alternative to IDF:

$$w_t^{DLITE} = DLITE(P_t, Q_t)$$

`Term weight based on the amount of DLITE in observing the term.`

Application: DLITE Alternative to BM25

Combined with TF, we have *iDL* (*ideal*):

1. $iDL_{dt} = w_{dt}^{TF} \times w_t^{DLITE}$
2. $iDL_{dt}^{\frac{1}{3}} = w_{dt}^{TF} \times \sqrt[3]{w_t^{DLITE}}$

Compared to Okapi BM25:

- $BM25 = w_{dt}^{TF} \times w_t^{IDF}$

Experimental Setup

Ad hoc information retrieval experiments:

- Lucene with a highly regarded BM25 baseline implementation.
- 3 benchmark collections: 1992 - 2017.
- Evaluation metrics: gMAP, MAP, P_{10}, nDCG, R_{PR}

Results: Best on Each Collection

Method	gMAP	MAP	P10	nDCG	R_{PR}
TREC 1994 Routing Track					
$BM25$	0.288	0.407	0.597	0.504	0.451
iDL	0.305	0.414	**0.639**	0.509	0.467
$iDL^{\frac{1}{3}}$	**0.309**	**0.419**	0.637	**0.524**	**0.469**
TREC 2005 HARD Track					
$BM25$	0.271	0.337	0.509	0.387	0.371
iDL	0.306	0.369	0.533	0.412	0.421
$iDL^{\frac{1}{3}}$	**0.323**	**0.388**	**0.564**	**0.447**	**0.447**
TREC 2017 Common Core					
$BM25$	0.387	0.457	0.615	0.452	0.452
iDL	**0.394**	0.468	**0.642**	**0.477**	**0.468**
$iDL^{\frac{1}{3}}$	0.387	**0.470**	0.612	0.465	0.424

TABLE I

BEST RESULTS ON EACH COLLECTION. EACH SCORE IS THE HIGHEST A METHOD ACHIEVED IN THE GIVEN EVALUATION METRIC. A BOLD FONT SHOWS THE BEST AMONG THE THREE METHODS IN EACH METRIC.

Results: HARD Track

Method	gMAP	MAP	P10	nDCG	R_{PR}
No Stemming					
$BM25$	0.271	0.337	0.509	0.357	0.371
iDL	0.306	0.369	0.533	0.410	0.421
$iDL^{\frac{1}{3}}$	**0.323**	**0.388**	**0.565**	**0.446**	**0.447**
With Stemming					
$BM25$	0.252	0.324	0.491	0.387	0.371
iDL	0.287	0.358	0.532	0.412	0.403
$iDL^{\frac{1}{3}}$	**0.312**	**0.382**	**0.548**	**0.447**	**0.436**

TABLE VII
TREC'05 HARD W. QUERY TITLE+DESC+NARR

Results: TREC'17 Common Core

Method	gMAP	MAP	P10	nDCG	R_{PR}
No Stemming					
$BM25$	0.353	0.439	0.591	0.472	0.466
iDL	**0.377**	0.449	**0.624**	**0.478**	0.472
$iDL^{\frac{1}{3}}$	0.320	**0.455**	0.612	0.465	**0.474**
With Stemming					
$BM25$	0.375	0.450	0.615	0.452	0.486
iDL	**0.392**	**0.464**	**0.642**	**0.468**	**0.497**
$iDL^{\frac{1}{3}}$	0.386	0.456	0.596	0.446	0.477

TABLE IX

TREC'17 COMMON CORE WITH QUERY TITLE+DESC

Findings

Findings:

- Experiments showed superior results with DLITE methods
- They consistently outperformed BM25, a very competitive baseline
- Results were even better on verbose (longer) queries and HARD topics.

Conclusion

Conclusion and looking forward:

- DLITE theory suitable for term weighting and applicable in other tasks for text mining and analytics.
- Can be applied in other processes to measure information gain in machine learning models, e.g. decision tree building.
  - $\sqrt[3]{DLITE}$ is a metric **distance** (quantity)
  - $DLITE$ can be considered **volumetric** (amount)
- Other creative ideas and research collaboration...

Thank you!

- Questions or comments?
- Feel free to reach out to me.

Weimao Ke
wk@drexel.edu