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

Weimao Ke
College of Computing and Informatics
Drexel University
Philadelphia, U.S.A.
wk@drexel.edu

2022 IEEE International Conference on Big Data (Big Data) | 978-1-6654-8045-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/BigData55660.2022.10020937

***Abstract*—The IDF (Inverse Document Frequency) term weighting method is a classic treatment of a term's significance in information retrieval and text analytics. IDF can be derived from the information-theoretic Kullback-Leibler (KL) Divergence and has given rise to competitive methods such as TF*IDF and Okapi BM25, which is the default scoring function of ElasticSearch. We developed a new information metric called DLITE and derived from it an alternative to IDF, namely iDL, for term weighting and scoring in ranked information retrieval. In a series of experiments we conducted on multiple benchmark Text REtrieval Conference (TREC) collections, iDL methods consistently outperformed BM25, a very competitive baseline, for ad hoc retrieval. We outline the theoretical properties of DLITE that support the effectiveness of iDL. As a general information measure, we expect DLITE to be applicable in many other areas of big-data analytics and machine learning where further research will be valuable.**

***Index Terms*—term weighting, information retrieval, probability distribution, entropy, information measure, relevance, ranking, effectiveness, metric**

I. Introduction

Research has relied on Shannon's information entropy and its derivatives in a wide range of processes for information retrieval (IR) and data mining (DM) [30], [32]. Information and probability theories provide essential guidance to the development of probabilistic retrieval and language modeling [27].

In particular, Kullback-Leibler (KL) divergence (relative entropy) and mutual information lay a robust theoretical foundation for the term significance measure of Inverse Document Frequency (IDF) [2], [5], [21]. IDF quantifies the amount of information of observing a term in a specific document by measuring its KL divergence from the collection-wide probability distribution.

In ranked retrieval, we often need to aggregate scores such as those based on TF*IDF and BM25, which include a similar component of IDF [28]. This practice presumes related weighting functions to be additive and bounded. However, such properties are not necessarily present in their theoretical counterparts, e.g., in KL divergence. Besides, research sometimes requires metric distances in related processes, where triangular inequality is desirable.

Inspired by the success of IDF and its limits, we have identified a new information measure, namely DLITE, and derived from it an alternative term weighting method we refer to as iDL. In this study, we apply iDL to ranked retrieval and compare it to BM25 on multiple benchmark Text REtrieval Conference (TREC) datasets. Results show superior performances of iDL compared to the classic, competitive baseline.

II. Related Works

Term frequency (TF) and document frequency (DF) are important statistics for term weighting in information retrieval and text mining. While term frequency (TF) exhibits the degree of a document's association with a term, inverse document frequency (IDF) is a manifestation of a term's specificity and discriminative power, a crucial indicator of document relevance [35].

While a term's IDF is equivalent to the mutual information between the term and the collection [33], the probabilistic retrieval framework provides a probabilistic interpretation of IDF weights as well [26]. Mutual information is equivalent to *relative entropy* that quantifies the difference between the joint probabilities and product probabilities of two random variables [12].

A. IDF as KL Divergence

IDF can also be regarded as Kullback-Leibler (KL) divergence of a term's occurrence in a document from its probability distribution in the entire collection [1]. Given a collection of N documents, the probability of drawing a document containing term t can be estimated by:

$$q_t = \frac{n_t}{N} \tag{1}$$

where n_t is the number of documents containing t. The complementary probability of drawing a document without term t is $q'_t = \frac{N-n_t}{N}$. For a specific document that contains the term, the probability of the term's occurrence is certain, i.e. $p_t = 1$ and $p'_t = 0$.

This research was partially supported by the LEADING project (IMLS grant #RE-246450-OLS-20) and NSF grant #1741306. Kashti Usmare contributed to the proof of DLITE's triangular inequality properties.

Given the definition of KL divergence for distributions P and Q:

$$KL(P||Q) = \sum_{x \in X} p_x \log \frac{p_x}{q_x} \tag{2}$$

We can compute KL divergence of term t:

$$\begin{aligned} w_t &= KL(P_t||Q_t) \qquad (3)\\ &= p_t \log \frac{p_t}{q_t} + p'_t \log \frac{p'_t}{q'_t} \qquad (4)\\ &= 1 \times \log \frac{1}{\frac{n_t}{N}} + 0 \times \log \frac{p_t}{q_t} \qquad (5)\\ &= \log \frac{N}{n_t} \qquad (6)\end{aligned}$$

w_t is exactly the classic IDF weight of term t. This reflects the amount of KL divergence in the term's occurrence (certainty) measured from the distribution obtained under a random process (collection-wide distribution) [6].

KL divergence (relative entropy) measures information for discrimination between two probability distributions by quantifying the entropy change in a non-symmetric manner [22]. Its values are unbounded and do not satisfy triangular inequality. Research has also employed KL information in language modeling to measure the difference between a document and query models for ranking and demonstrated strong performances [23], [36].

B. IDF in TF*IDF

Let tf_{dt} denote the term frequency of t in document d. Its TF*IDF weight can be computed by:

$$TFIDF_{dt} = tf_{dt} \times \log \frac{N}{n_t} \tag{7}$$

There are variations of TF*IDF where TF is log-transformed or normalized by document length:

$$\begin{aligned} TFIDF^{log}_{dt} &= (1 + \log tf_{dt}) \times \log \frac{N}{n_t} \qquad (8)\\ TFIDF^{norm}_{dt} &= \frac{tf_{dt}}{l_d} \times \log \frac{N}{n_t} \qquad (9)\end{aligned}$$

where l_d is the length of document d, i.e., the number of terms it contains. Document length normalization reduces term frequencies to probability estimates and is a common practice in ranked retrieval.

C. IDF in BM25

Another classic variation of TFIDF is Okapi BM25, which is derived from probabilistic models [28]. In BM25, the TF weight is normalized by document length as well as a saturation function:

$$w^{TF}_{dt} = \frac{tf_{dt}}{tf_{dt} + k\big((1-b) + b\frac{l_d}{avl}\big)} \tag{10}$$

where avl is the average document length in the collection. Whereas b controls the degree of document length normalization (0 for no normalization and 1 for full-scale normalization), k is the pivot value that determines how quickly an increasing score saturates.

BM25 also includes an IDF component with a smoothed probability estimate:

$$w^{IDF}_t = \log \frac{N - n_t + 0.5}{n_t + 0.5} \tag{11}$$

The final BM25 weight is the product of w^{TF}_{dt} and w^{IDF}_t:

$$\begin{aligned} BM25_{dt} &= w^{TF}_{dt} \times w^{IDF}_t \qquad (12)\\ &= \frac{tf_{dt}}{tf_{dt} + k\big((1-b) + b\frac{l_d}{avl}\big)} \qquad (13)\\ &\quad \times \log \frac{N - n_t + 0.5}{n_t + 0.5} \qquad (14)\end{aligned}$$

Experiments have shown competitive results based on $0.5 < b < 0.8$ and $1.2 < k < 2$. For TREC ad hoc retrieval, years of evaluation results indicate that best results can be achieved by roughly $b = 0.75$ and $k = 1.5$.

III. DLITE Theory

The Discounted Least Information Theory of Entropy (DLITE) is an extension of our prior work on the Least Information Theory (LIT) that satisfies several additional properties as an information metric. We shall introduce the LIT measure first.

A. LIT Measure

The Least Information Theory (LIT) quantifies the amount of entropic difference between two probability distributions [8], [17]. Given distributions P and Q of variable X, LIT is computed by:

$$\begin{aligned} LIT(P,Q) &= \sum_{x \in X} \int_{p_x}^{q_x} -\log p \; dp \qquad (15)\\ &= \sum_{x \in X} \Big| p_x(1 - \ln p_x) - q_x(1 - \ln q_x) \Big| (16)\end{aligned}$$

where x is one of the mutually exclusive inferences of X, and p_x and q_x are probabilities of x on the P and Q distributions respectively.

For any probabilities p and q, let:

$$g(p,q) = \Big| p(1 - \ln p) - q(1 - \ln q) \Big| \tag{17}$$

LIT can be written as:

$$LIT(P,Q) = \sum_{x \in X} g(p_x, q_x) \tag{18}$$

Research has applied LIT to data clustering, classification, and information retrieval, and shown its competitive performances compared to classic baselines [10], [13], [17]–[19].

B. Entropy Discount

For DLITE, we introduce the following entropy discount:

$$\Delta_H(P,Q) = \sum_{x\in X} \Big|p_x - q_x\Big| \frac{\int_{p_x}^{q_x} -p\log p \ dp}{\int_{p_x}^{q_x} x \ dx} \tag{19}$$

$$= \sum_{x\in X} \frac{\Big|p_x^2(1-2\ln p_x) - q_x^2(1-2\ln q_x)\Big|}{2(p_x+q_x)} \tag{20}$$

For any probabilities p and q, let:

$$\delta_h(p,q) = \frac{\Big|p^2(1-2\ln p) - q^2(1-2\ln q)\Big|}{2(p+q)} \tag{21}$$

The entropy discount Δ_H can be written as:

$$\Delta_H(P,Q) = \sum_{x\in X} \delta_h(p_x, q_x) \tag{22}$$

C. DLITE: LIT with Entropy Discount

We now define the Discounted Least Information Theory of Entropy, or DLITE, as the amount of LIT subtracted by its entropy discount Δ_H:

$$DL(P,Q) = LIT(P,Q) - \Delta_H(P,Q) \tag{23}$$

$$= \sum_{x\in X} g(p_x,q_x) - \delta_h(p_x,q_x) \tag{24}$$

For any probability change from p to q, let:

$$dl(p,q) = g(p,q) - \delta_h(p,q) \tag{25}$$

Equation 24 can written as:

$$DL(P,Q) = \sum_{x\in X} dl(p_x,q_x) \tag{26}$$

D. DLITE's Information-theoretic Properties

Again, DLITE is the amount of LIT with the Δ_H discount:

$$DL(P,Q) = LIT(P,Q) - \Delta_H(P,Q) \tag{27}$$

$$= \sum_{x\in X} \int_{p_x}^{q_x} \log \frac{1}{p} \ dp \tag{28}$$

$$- \sum_{x\in X} \Big|p_x - q_x\Big| \frac{\int_{p_x}^{q_x} p \log \frac{1}{p} \ dp}{\int_{p_x}^{q_x} p \ dp} \tag{29}$$

Whereas LIT represents the sum of weighted, microscopic entropy changes, it consists of an amount of entropy change due to the scale of related probabilities, leading to an undesirable consequence of having different LIT amounts in different sub-system breakdowns. The entropy discount, Δ_H, accounts for this extra amount in the LIT and reduces it to a scale-free measure. As shown in Equation 29, the discount on each x dimension is a product of the absolute probability change in p and a mean of $\log \frac{1}{p}$.

We discussed justifications of DLITE with a list of metric and information-theoretical properties in [20]. The latest update on mathematical proofs, including those on its triangular inequality properties, can be found in the technical report [37]. We highlight DLITE's major theoretical properties below.

Research has proposed many basic properties desirable of any information measure or a coefficient of divergence. The DLITE measure meets several of these properties, such as those listed in [3], [21]. Here we list major DLITE characteristics as an information quantity:

1) DLITE is defined for all pairs of probabilities on two distributions P and Q.
2) $DL(P,Q)$ increases when absolute pairwise probability differences of P and Q increase.
3) DLITE increases with an increasing number of equiprobable inferences, when reducing that uniform distribution to a certainty.
4) DLITE is non-negative, with its maximum at 1.
5) DLITE of an ensemble (entire system) is the weighted sum of DLITEs in its sub-systems.
6) It is additive for independent variables as well as dependent probability distributions in special cases.
7) The sum of DLITE distances $\sqrt[3]{DL}$ on two variables is no less than $\sqrt[3]{DL}$ on their product distributions.

To demonstrate some of these properties, Figure 1 compares DLITE to classic information measures including Shannon Entropy [31], KL divergence [21], and Jensen-Shannon (JS) Divergence [24]. DLITE is bounded in $[0,1]$ regardless of the dimensionality. The quantity on one single inference $x \in X$ is maximized, $dl(p_x,q_x) = 0.5$, when the probability changes from $p_x = 0$ to $q_x = 1$, or from $p_x = 1$ to $q_x = 0$. With 2 mutually exclusive inferences, the overall DLITE is maximized for changes from $P = (0,1)$ to $Q = (1,0)$, where $DL(P,Q) = 1$.

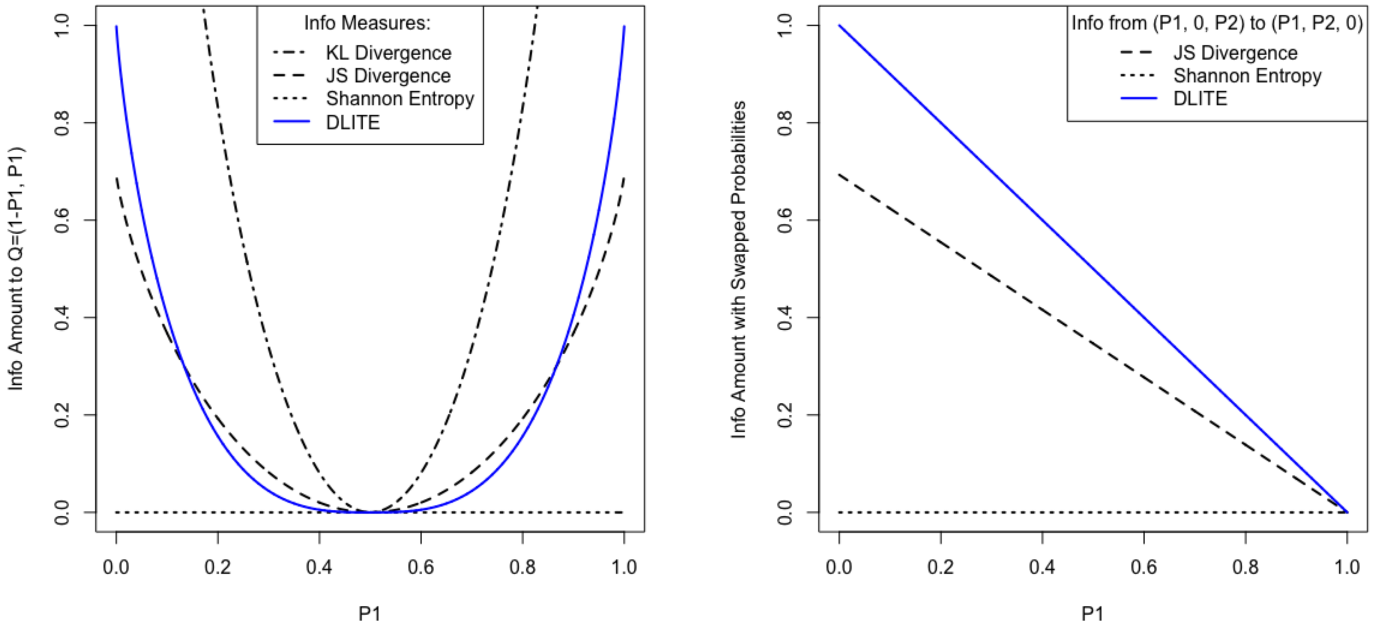
Shannon entropy, on the other hand, always returns 0 for swapped probabilities, as shown by examples in Figures 1 (a) and (b). In Figure 1 (a), KL Divergence approaches infinity with a 0 probability whereas DLITE and JS divergence are bounded by 1 and $\ln 2$ respectively. Likewise, as Figure 1 (b) shows, DLITE is bounded in $[0,1]$, when 2 out of 3 probabilities are swapped.

E. DLITE's Metric Properties

Given the definition in Equation 24 or 29, it can be shown that DLITE satisfies the following metric properties:

1) Non-negativity: $DL(P,Q) \geq 0$ for any probability distributions P and Q of the same dimensionality.
2) Identity of Indiscernibles: $DL(P,Q) = 0$ if and only if P and Q are identical distributions.
3) Symmetry: $DL(P,Q) == DL(Q,P)$, the amount of the information from P to Q is the same as that from Q to P.

Figure 2 plots the value of DLITE, $dl(p,q)$, for any probability change from p to q and demonstrates the above three



(a) Binary swap $P(p_1, p_2) \rightarrow Q(p_2, p_1)$

(a) Swap 2 of 3 probabilities

Fig. 1. Information for Swapped Probabilities. X axis denotes the probability of one inference whereas Y shows the amount of information, with (a) swapped probabilities in the binary case and (b) swapped probabilities of 3 inferences. Compare to Fig. 1 in [24].

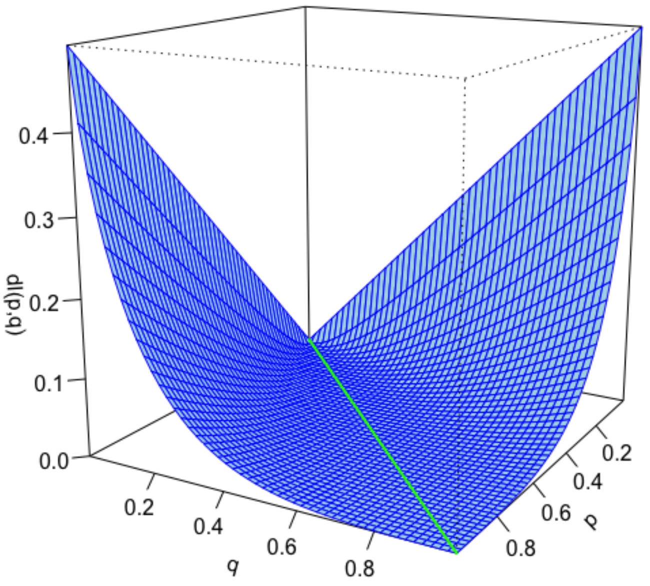


Fig. 2. $dl(p, q)$ for any p and q values

properties: (1) all values ≥ 0, (2) 0 values only on the diagonal line (in a lighter color) where $p = q$, and (3) a symmetry indicating $dl(p, q) = dl(q, p)$.

We observe that DLITE's cube root $\sqrt[3]{dl}$ satisfies triangular inequality:

$$\sqrt[3]{dl(p,q)} + \sqrt[3]{dl(q,r)} \geq \sqrt[3]{dl(p,r)} \tag{30}$$

Applying Minkowski's inequality [14] to the above, we also prove that $\sqrt[3]{DL}$ satisfies the triangular inequality:

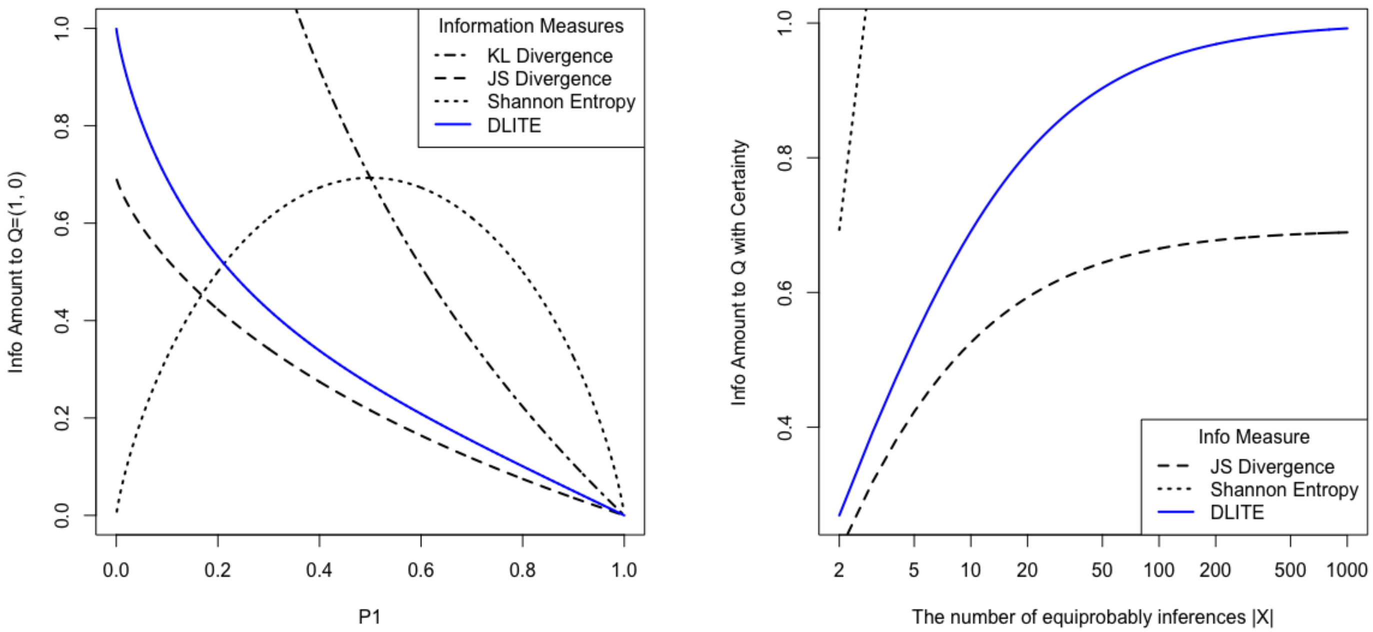
$$\sqrt[3]{DL(P,Q)} + \sqrt[3]{DL(Q,R)} \geq \sqrt[3]{DL(P,R)} \tag{31}$$

where P, Q, and R are probability distributions of the same dimensionality.

Given the above properties of DLITE, it is straightforward to show that $\sqrt[3]{DL}$ also satisfies non-negativity, the identity of indiscernibles, and symmetry. Therefore, the cube root is a metric distance, and we can regard DLITE as a *volumetric* measure in the amount of information. We refer to $\sqrt[3]{DL}$ as the *DLITE distance*. Detailed proofs of these and related properties can be found in [20], [37].

We can compare this characteristic to that of Jessen-Shannon (JS) Divergence, of which the square root is a metric [11], [24]. DLITE and JS Divergence share similar patterns in the measured amount of information.

In Figure 3 we compare DLITE with classic measures on reducing a probability distribution to certainty (when one inference becomes the ultimate outcome). Figure 3 (a) compares the measures of reducing a binary probability distribution $P(p_1, p_2)$ to certainty $Q(1, 0)$, i.e. with the first inferences as the ultimate outcome. Shannon entropy is symmetric because it only accounts for the overall entropy reduction and disregards the amount of probability change in specific inferences



(a) $P(p_1, p_2) \to Q(1, 0)$

(b) Equiprobable P to certainty Q

Fig. 3. DLITE vs. classic information measures on reducing to certainty. Y is the amount of information $I(P, Q)$ based on each information measure. (a) is the binary case, where X denotes probability p_1 of two mutually exclusive inferences, with $p_2 = 1 - p_1$. (b) shows the general case of reducing an equiprobable distribution P to certainty Q, where log-transformed X denotes the number of equiprobable inferences.

(dimensions). DLITE and Jessen-Shannon divergence follow a similar pattern with a upper bound, whereas the KL divergence is unbounded.

In Figure 3 (b), we compare the information measures in reducing equiprobable inferences to certainty. With an increasing number of equiprobable inferences, Shannon entropy continues to increase, whereas DLITE and JS divergence are bounded. DLITE approaches 1 asymptotically as the number of equiprobable inferences approaches infinity.

F. Implications of DLITE Properties

The above DLITe properties have important implications in applications such as Information Retrieval and Text Mining. We highlight some of these properties and how they can make a difference compared to classic treatments:

- Given the metric properties of DLITE, it is fitting to use it as a distance measure where the sum of such values (scores) are meaningful. This is compatible with classic IR ranking methods where the scoring function is the sum of matching term weights.
- Because DLITE is bounded in $[0, 1]$, it is different from an unbounded function such as the KL divergence. IDF as a direct application of KL divergence can have infinite (theoretically) or extremely large values (practically) that dominate the scoring function. As a result, when the query contains a rare term, that term's large IDF will render other terms useless for retrieval ranking. DLITE does not inherit this problem thanks to the upper bound.
- As the DLITE of an ensemble (the entire system) can be computed as the weighted sum of its sub-systems, it offers researchers the capacity to dissect and reconstruct a system in different ways without altering the final scoring.

IV. iDL Term Weighting

In this work, we apply the DLITE theory to ad hoc information retrieval (IR), particularly for term weighting. In the bag-of-words approach to IR, we view a document as a set of terms with probabilities (estimated by frequencies) of occurrences. By analyzing a term's probability (frequency) in a document vs. that in the collection, we can compute the amount of information in the term's occurrence to weight the term. We conjecture that the greater amount of DLITE a term has, the more heavily the term should be weighted to represent the document.

A. DLITE Alternative to IDF

Following the discussed KL divergence model for IDF, we compute a term's weight based on its DLITE from its probability distribution in the entire collection [1]. Again, given a collection of N documents, the probability of drawing a document containing term t can be estimated by:

$$q_t = \frac{n_t}{N} \tag{32}$$

where n_t is the number of documents containing t. The complementary probability of drawing a document without term t is $q_t' = \frac{N-n_t}{N}$. For a specific document that contains the term, the probability of the term's occurrence is certain, i.e. $p_t = 1$ and $p_t' = 0$.

Based on the DLITE definition, we compute the weight of term t based on P_t (document) and Q_t (collection) distributions:

$$w_t^{DLITE} = DLITE(P_t, Q_t) \tag{33}$$

$$= LIT(P_t, Q_t) - \Delta_H(P_t, Q_t) \tag{34}$$

$$\begin{aligned} = & \int_0^{q_t'} \log \frac{1}{p}\, dp + \int_{q_t}^{1} \log \frac{1}{p}\, dp \\ & - q_t' \frac{\int_0^{q_t'} p \log \frac{1}{p}\, dp}{\int_0^{q_t'} p\, dp} \\ & - (1 - q_t) \frac{\int_{q_t}^{1} p \log \frac{1}{p}\, dp}{\int_{q_t}^{1} p\, dp} \end{aligned} \tag{35}$$

$$\begin{aligned} = & q_t'(1 - \ln q_t') - \frac{1}{2} q_t'(1 - 2 \ln q_t') \\ & + (1 - q_t(1 - \ln q_t)) \\ & - \frac{1 - q_t^2(1 - 2 \ln q_t)}{2 + 2q_t} \end{aligned} \tag{36}$$

$$\begin{aligned} = & \frac{q_t'}{2} + (1 - q_t(1 - \ln q_t)) \\ & - \frac{1 - q_t^2(1 - 2 \ln q_t)}{2 + 2q_t} \end{aligned} \tag{37}$$

where $q_t = \frac{n_t}{N}$ and $q_t' = \frac{N-n_t}{N}$.

B. iDL with TF

We adopt the same TF component as in BM25:

$$w_{dt}^{TF} = \frac{tf_{dt}}{tf_{dt} + k\big((1-b) + b\frac{l_d}{avl}\big)} \tag{38}$$

where avl is the average document length, b is the parameter for document length normalization, and k is the saturation pivot value.

Finally, we conbine w_t^{DLITE} and w_{dt}^{TF} using their product:

$$iDL_{dt} = w_{dt}^{TF} \times w_t^{DLITE} \tag{39}$$

C. $iDL^{\frac{1}{3}}$ with TF

We showed that $\sqrt[3]{DLITE}$ meets triangular inequality and is a metric distance. Here we propose a second term weighting method based on the product of w_{dt}^{TF} and $\sqrt[3]{w_t^{DLITE}}$:

$$iDL_{dt}^{\frac{1}{3}} = w_{dt}^{TF} \times \sqrt[3]{w_t^{DLITE}} \tag{40}$$

D. A Note on Computational Complexity

As shown in the equations above, iDL_{dt} and $iDL_{dt}^{\frac{1}{3}}$ scores depend on q_t, which is estimated by the term's document frequency (DF) n_t, whereas w_{dt}^{TF} is a function of its term frequency in the document (TF). Same as TF*IDF and BM25, both iDL and $iDL^{\frac{1}{3}}$ are a function of TF and DF statistics. They are all based on a logarithmic transformation with normalization of these statistics. In short, the alternative iDL methods proposed in this research does not incur additional computational costs than classic BM25 or TF*IDF does.

V. Experimental Setup

A. Data Collections and Topics

We used the following Text REtrieval Conference (TREC) benchmark datasets, from the Linguistic Data Consortium and NIST, for retrieval experiments: TREC 1994 Vol 2 Disk 3, TREC 2005 Hard (High Accuracy Retrieval from Documents) track, and TREC 2017 Common Core track. These are representative TREC datasets of three decades and have been widely used for ad hoc retrieval experiments. We use the following topics (queries) and relevance judgment provided by TREC in each year:

- TREC 2 routing topics 51 - 100 with title, description, summary, narrative, and concepts (disk 3) [29];
- TREC 2005 HARD/Robust 50 topics (303 - 689) with title, description, and narrative for retrieving the AQUAINT I data [38].
- TREC 2017 Common Core topics (303 - 690) with title, description, and narrative reused for experiments on the New York Times Annotated Corpus [4].

These collections represent a diversity of text data and query tasks. In TREC 2, for example, the *concepts* field in 51 - 100 topics contains a verbose list of concepts to represent each search topic. Text queries automatically generated from the concept lists are likely to be more accurate than general descriptions in sentences. TREC 2005 HARD topics were developed as a list of *difficult* topics from previous years' ad hoc experiments. TREC 2017 reused and revised queries from the 2004 Robust track to bring them up to date. These were run and re-evaluated based on a more recent collection of 1.8 million NYT articles. Using these diverse data and topics enabled a relatively thorough examination of the proposed methods' effectiveness in various domain and task contexts.

B. Experimental System

We implemented the retrieval ranking methods using the Lucene core search engine library in Java [15]. We reused the Okapi BM25 implementation reported in [25] and validated by [27], which achieved highly competitive results in TREC. We set parameter values $b = 0.75$ and $k_1 = 1.5$ for BM25, according to reported best results on related data. We used the same b and k parameters for the TF component in iDL.

We performed standard tokenization, case-folding, and stopword removal before indexing documents. For each data collection, we conducted one set of experiments with stemming and the other without it. We also used different query verbosity levels in the searches, with query title, description, or narrative.

C. *Evaluation Metrics*

We used human relevance judgment (QRELs) NIST developed for TREC 2, TREC 2005 HARD tracks, and TREC 2017 common core track as the ground truth for retrieval evaluation. We compared the proposed methods based on DLITE, namely iDL and $iDL^{\frac{1}{3}}$, to Okapi BM25, a very strong baseline.

Evaluation metrics included mean average precision with arithmetic averaging (MAP) and geometric (gMAP), best precision at rank 10, normalized discounted cumulative gain at 10 ($nDCG_{10}$), and recall precision. While arithmetic average MAP provides a simple mean score across multiple queries, the geometric average (gMAP) is sensitive to poorly performed tasks and is a very useful metric developed for the 2005 HARD track [38]. NDCG favors early retrieval of highly relevant documents in a ranked list and has become widely adopted for ranked retrieval evaluation [16].

VI. Experimental Results

Overall, the proposed DLITE methods overwhelmingly outperformed BM25 in experiments. Table I is a summary of best results achieved by each method, per evaluation metrics on each data collection. As shown in Table I, iDL and $iDL^{\frac{1}{3}}$ dominate the best results in every evaluation metric. The results are consistent across the benchmark data collections developed in a period of more than 20 years (1994 - 2017).

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.

We elaborate on each set of experiments in sections VI-A - VI-C below. In each of Tables II - IX, we report on one set of experiments conducted with stemming and the other without. We highlight the best scores in each evaluation metric in **bold** fonts.

A. *TREC 1994 Routing on Vol 2 Disk 3*

Table II shows results from experiments using query terms from topic titles. Without stemming, $iDL^{\frac{1}{3}}$ performed best in terms of every evaluation metric. Stemming further improved the results for DLITE methods, where $iDL^{\frac{1}{3}}$ continued to dominate the best results (in MAP, nDCG, and R_{PR}). iDL also consitently outperformed BM25.

Method	gMAP	MAP	P10	nDCG	R_{PR}
No Stemming					
$BM25$	0.187	0.334	0.495	0.362	0.361
iDL	0.195	0.354	0.497	0.374	0.384
$iDL^{\frac{1}{3}}$	**0.203**	**0.360**	**0.495**	**0.380**	**0.392**
With Stemming					
$BM25$	0.183	0.329	0.440	0.338	0.353
iDL	**0.202**	0.370	**0.489**	0.396	0.398
$iDL^{\frac{1}{3}}$	0.174	**0.376**	0.483	**0.402**	**0.407**

TABLE II

TREC 1994 with Query Title (Disk3)

Method	gMAP	MAP	P10	nDCG	R_{PR}
No Stemming					
$BM25$	0.206	0.356	0.503	0.398	0.387
iDL	**0.224**	0.384	**0.579**	0.431	0.422
$iDL^{\frac{1}{3}}$	0.201	**0.393**	0.560	**0.432**	**0.427**
With Stemming					
$BM25$	0.220	0.335	0.460	0.354	0.362
iDL	**0.246**	0.374	0.533	0.414	0.405
$iDL^{\frac{1}{3}}$	0.229	**0.394**	**0.544**	**0.431**	**0.424**

TABLE III

TREC 1994 with Query Title+Desc

Table III continues to show superior performance of DLITE methods over BM25, using query titles and descriptions.

Method	gMAP	MAP	P10	nDCG	R_{PR}
No Stemming					
$BM25$	0.288	0.407	0.597	0.504	0.451
iDL	**0.302**	**0.410**	**0.639**	**0.509**	**0.456**
$iDL^{\frac{1}{3}}$	0.299	0.408	0.634	0.493	0.453
With Stemming					
$BM25$	0.281	0.399	0.565	0.488	0.442
iDL	0.305	0.414	0.623	0.523	0.467
$iDL^{\frac{1}{3}}$	**0.309**	**0.419**	**0.637**	**0.524**	**0.469**

TABLE IV

TREC 1994 with Query Concepts (Disk3)

In TREC 2 topics, each query also comes with a verbose list of concepts (accurate keywords). With these manually picked concept terms, which are overall quite precise in defining each topic, Table IV shows further improvement of DLITE methods – whereas iDL produced the best results without stemming in all metrics, $iDL^{\frac{1}{3}}$ did that with stemming.

B. *TREC 2005 HARD Track*

In TREC 2005 HARD/Robust track, the 50 topics were considered difficult retrieval tasks. We used title, description, and title+description as queries in the experiments. As Table V shows, $iDL^{\frac{1}{3}}$ is the best method for searches based on titles with and without stemming, in terms of all evaluation metric. iDL outperformed BM25 as well.

When we used topic title and *descriptions* for query representation, as shown in Table VI, $iDL^{\frac{1}{3}}$ and iDL continued to outperform BM25.

Method	gMAP	MAP	P10	nDCG	R_{PR}
No Stemming					
$BM25$	0.169	0.274	0.403	0.256	0.298
iDL	0.184	0.301	0.408	0.316	0.345
$iDL^{\frac{1}{3}}$	**0.189**	**0.308**	**0.423**	**0.333**	**0.355**
With Stemming					
$BM25$	0.162	0.261	0.377	0.270	0.293
iDL	0.173	0.277	0.420	0.317	0.318
$iDL^{\frac{1}{3}}$	**0.183**	**0.295**	**0.441**	**0.348**	**0.341**

TABLE V
TREC'05 WITH QUERY TITLE (HARD)

Method	gMAP	MAP	P10	nDCG	R_{PR}
No Stemming					
$BM25$	0.221	0.293	0.454	0.322	0.334
iDL	0.241	0.322	0.437	0.336	0.370
$iDL^{\frac{1}{3}}$	**0.252**	**0.335**	**0.454**	**0.367**	**0.386**
With Stemming					
$BM25$	0.214	0.289	0.452	0.355	0.331
iDL	0.237	0.315	0.464	0.387	0.368
$iDL^{\frac{1}{3}}$	**0.261**	**0.339**	**0.502**	**0.411**	**0.399**

TABLE VI
TREC'05 HARD WITH QUERY TITLE+DESC

We observed the same consistent results using query title, description, and narratives. DLITE methods' evaluation scores have a higher margin over those of BM25.

TREC 2005 HARD topics represent *difficult* information needs, for which query specification is challenging. The proposed methods appeared to perform better with these *challenging* tasks, as was so suggested by the higher gMAP scores in the experiments.

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

Method	gMAP	MAP	P10	nDCG	R_{PR}
No Stemming					
$BM25$	0.202	0.388	0.459	0.357	0.410
iDL	0.207	**0.389**	0.479	0.358	0.413
$iDL^{\frac{1}{3}}$	**0.208**	0.389	**0.480**	**0.366**	**0.413**
With Stemming					
$BM25$	0.222	**0.412**	**0.527**	**0.390**	**0.443**
iDL	**0.224**	0.405	0.501	0.385	0.437
$iDL^{\frac{1}{3}}$	0.223	0.403	0.499	0.387	0.434

TABLE VIII
TREC'17 COMMON CORE WITH QUERY TITLE

C. TREC 2017 Common Core

Using query titles on the recent TREC 2017 common core dataset, experiments produced mixed results with very close scores. While DLITE methods performed slightly better than BM25 without stemming, BM25 with stemming produced the best results in terms of metrics such as P_{10} and nDCG, as shown in Table VIII.

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

When query descriptions are also included in the searches, experiments showed significant improvements. As Table IX shows, DLITE methods, especially iDL, outperformed BM25 in all experimental settings. We observe that increasing query verbosity improved retrieval performances and gave a greater boost to iDL methods.

We want to note that, while our results are competitive and comparable to those reported in TREC, it is not a fair comparison to put them side by side with ours. In TREC, many systems were fine-tuned, sometimes with additional data and manual input. Participants have reportedly used external resources such as WordNet and Wikipedia to obtain the best results. We have limited our methods in this study to use provided TREC data only and to demonstrate their effectiveness in standard settings without additional variables.

We conduct these experiments to understand the application of DLITE for text processing and information retrieval. While these initial results are impressive out of the box, we plan to fine tune BM25 and the proposed alternatives, perform significance tests with randomization, and conduct a meta-analysis of results from a wider range of related experiments in future research [34].

VII. CONCLUSION

In this paper, we proposed an alternative to the classic IDF term weighting scheme, namely iDL, based on a new DLITE information measure. In a series of experiments on benchmark TREC collections, iDL consistently outperformed Okapi BM25 – a very competitive baseline in the latest research and the default scoring function of ElasticSearch [9] – and showed exceptionally superior results with longer queries. Overall, stemming also improved the proposed methods' effectiveness.

Several fundamental properties of DLITE may have contributed to the effectiveness of the proposed methods. As noted, DLITE is bounded, satisfies conditions as a metric distance[1], and is additive. Unlike KL divergence, on which

[1] Note the cube root of DLITE satisfies the triangular inequality.

BM25-IDF is based, DLITE quantities are finite (no greater than 1). These properties enable related term weights to be reasonably compared and aggregated in the scoring process.

DLITE offers a new measure to quantify information in probability distributions. While it is possible to derive other novel methods for IR from DLITE, one can incorporate DLITE into an existing probabilistic framework such as Divergence from Randomness (DFR) [7]. DLITE can also be used in machine learning (ML) models where an *information gain* criterion or a loss function is critical, e.g. for building a decision tree. With demonstrated experimental results in this work, we expect DLITE to be applicable in many other applications of big-data analytics where further research will be valuable.

References

[1] A. Aizawa. The feature quantity: an information theoretic perspective of TFIDF-like measures. In *SIGIR'00*, pages 104–111, 2000.

[2] A. Aizawa. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45 – 65, 2003.

[3] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):131–142, 1966.

[4] J. Allan, D. Harman, E. Kanoulas, D. Li, C. V. Gysel, and E. M. Voorhees. TREC 2017 common core track overview. In E. M. Voorhees and A. Ellis, editors, *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017*, volume 500-324 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2017.

[5] G. Amati and C. J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, Oct. 2002.

[6] G. Amati and C. J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, Oct. 2002.

[7] G. Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, 20(4):357–389, October 2002.

[8] Anonymous. Anonymous. *Anonymous*, xx(2):xxx–xxx, 20xx.

[9] S. Connelly. Practical bm25 - part 2: The bm25 algorithm and its variables. https://www.elastic.co/blog/practical-bm25-part-2-the-bm25-algorithm-and-its-variables.

[10] Y. Du, J. Liu, W. Ke, and X. Gong. Hierarchy construction and text classification based on the relaxation strategy and least information model. *Expert Systems with Applications*, 100:157–164, 2018.

[11] D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860, Sept. 2006.

[12] R. M. Fano. *Transmission of Information: A Statistical Theory of Communication*. MIT Press, March 1961.

[13] X. Gong and W. Ke. Term weighting for interactive cluster labeling based on least information gain. In *ACM WSDM 2015 workshop on heterogeneous information access (HIA'15). Shanghai, China*, 2015.

[14] G. Hardy. *Inequalities*. Cambridge University Press, second edition, 1988.

[15] E. Hatcher, O. Gospodnetić, , and M. McCandless. *Lucene in Action*. Manning Publications, second edition edition, March 2010.

[16] K. Jarvelin and J. Kekalainen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[17] W. Ke. Information-theoretic term weighting schemes for document clustering. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '13, pages 143–152, New York, NY, USA, 2013. ACM.

[18] W. Ke. Information-theoretic term weighting schemes for document clustering and classification. *International Journal on Digital Libraries*, 16(2):145–159, 2015.

[19] W. Ke. Text retrieval based on least information measurement. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '17, page 125–132, New York, NY, USA, 2017. Association for Computing Machinery.

[20] W. Ke. Dlite: The discounted least information theory of entropy, 2020.

[21] S. Kullback. *Information Theory and Statistics*. Wiley, New York, 1959.

[22] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[23] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 111–119, 2001.

[24] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, Jan 1991.

[25] J. Pérez-Iglesias, J. R. Pérez-Agüera, V. Fresno, and Y. Z. Feinstein. Integrating the Probabilistic Models BM25/BM25F into Lucene. *CoRR*, abs/0911.5046, 2009.

[26] S. Robertson. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60:503–520, 2004.

[27] S. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieva*, 3(4):333–389, 2009.

[28] S. Robertson and H. Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

[29] S. J. M. H.-B. M. G. S. Robertson, S. Walker. Okapi at trec-2. In *The Second Text REtrieval Conference*, pages 21 – 34. NIST, 1993.

[30] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, July and October 1948.

[31] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948.

[32] D. Shaw and C. H. Davis. Entropy and information: A multidisciplinary overview. *Journal of the American Society for Information Science*, 34(1):67–74, 1983.

[33] M. Siegler and M. Witbrock. Improving the suitability of imperfect transcriptions for information retrieval from spoken documents. In *ICASSP'99*, pages 505–508. IEEE Press, 1999.

[34] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, page 623–632, New York, NY, USA, 2007. Association for Computing Machinery.

[35] K. Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 60:493–502, 2004.

[36] T. Tao and C. Zhai. An exploration of proximity measures in information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 295–302, 2007.

[37] K. Umare and W. Ke. On triangular inequality properties of the discounted information theory of entropy (dlite). Technical report, Drexel University, Philadelphia, PA, U.S.A., Technical Report, 2022.

[38] E. Voorhees. Overview of trec 2005. In *Text Retrieval Conference (TREC)*, 2005.