Inverse Document Frequency (IDF)

A Convergence of Ideas for Information Retrieval

Weimao Ke

Drexel University

Table of contents

1. Information Retrieval Background

2. A Heuristic Approach

3. A Probabilistic Approach

4. An Information-Theoretic Approach

5. Shannon Entropy

Information Retrieval Background

Term Weighting

Search query:

The Maid of Orléans

- What documents (pages) should be retrieved?
- Which ones should be ranked early (top)?
- What terms (words) are more important?

Background

Observation:

- **The**: a stop (too common) word and hardly informative
- **Maid**: an informative word, but still broad (common)
- **of**: again, an almost meaningless stop word
- **Orleans**: a specific and meaningful word (not as common)

A Heuristic Approach

A Heuristic Approach

Document Frequency (DF) n_t: $\#$ documents containing term t.

- Common (broad) terms – large DF – should be given less weight.
- Rare (specific) terms – small DF – should have more weight.

IDF: A Heuristic Approach (Salton and McGill, 1986)

Given the inverse relation between DF (n_t) and expected weight w_t:

- Term weight w_t should be the inverse of n_t
- 1^{st} proposal: $w_t = \frac{1}{n_t}$
- 2^{nd} proposal with N: $w_t = \frac{N}{n_t}$
- 3^{rd} proposal with logarithm[1]: $w_t = \log \frac{N}{n_t}$

N is the number of documents, $N > n_t$.

[1] Logarithm examples: $\log_{10} 1 = 0$, $\log_{10} 10 = 1$, and $\log_{10} 100 = 2$.

A Probabilistic Approach

IDF: A Probabilistic Approach (Robertson and Zaragoza, 2009)

The probability of a document d being relevant to query q, based on Bayesian Theorem:

$$P(Relevant|d) = \frac{P(d|Relevant)P(Relevant)}{P(d)} \tag{1}$$

IDF: A Probabilistic Approach (Robertson and Zaragoza, 2009)

Equivalent to log-likelihood of odds[2]:

$$\sum_{t \in q} \log \frac{P(t|Relevant)[1 - P(t|NonRelevant)]}{[1 - P(t|NonRelevant)]P(t|NonRelevant)} \tag{2}$$

$$\approx \sum_{t \in q} \log \frac{(r_t + 1)(N - R - n_t + r_t + 1)}{(n_t - r_t + 1)(R - r_t + 1)} \tag{3}$$

where:

- N is the total number of documents;
- n_t is the number of documents containing term t (DF);
- R is the total number of relevant documents;
- r_t is the number of *relevant documents* containing t;

[2] +1 in the formula to avoid zero probabilities.

IDF: A Probabilistic Approach (Robertson and Zaragoza, 2009)

1. Given a query, the number of relevant docs R is fixed (constant);
2. In reality, we do NOT know what relevant documents are and assume a fixed r_t for all terms;

In this case, $r_t + 1$ and $R - r_t + 1$ are constant and can be ignored.

IDF: A Probabilistic Approach (Robertson and Zaragoza, 2009)

In the end, we have:

$$\sum_{t \in q} \log \frac{N - R - n_t + r_t + 1}{n_t - r_t + 1} \tag{4}$$

If relevant documents are a small subset of all documents, $N \gg R$ and $n_t \gg r_t$. The above becomes:

$$\approx \sum_{t \in q} \log \frac{N - n_t + 1}{n_t + 1} \tag{5}$$

IDF: A Probabilistic Approach (Robertson and Zaragoza, 2009)

For infrequent terms $N \gg n_t \gg 1$:

$$\approx \sum_{t \in q} \log \frac{N}{n_t} \tag{6}$$

which is the sum of IDF weights for terms in the query.

An Information-Theoretic Approach

IDF: An Information-Theoretic Approach

Information Theory and Development:

- Shannon (1948): *entropy* as missing information
- Kullback and Leibler (1951): *KL divergence* or relative entropy
- An important application of KL is *mutual information*

IDF: An Information-Theoretic Approach

KL divergence is defined as:

$$KL(P||Q) = \sum_{x \in X} p_x \log \frac{p_x}{q_x} \tag{7}$$

where P and Q are the (true) probability and estimate distributions of the same variable X.

IDF: An Information-Theoretic Approach

Given a collection of N documents: If you randomly draw a document, how likely does it contain term t?

$$q_t = \frac{n_t}{N} \tag{8}$$

$$q'_t = \frac{N - n_t}{N} \tag{9}$$

n_t is the number of documents containing t.

q'_t denotes the probability that term t does NOT appear.

IDF: An Information-Theoretic Approach

Now, for ONE document that contains the term, it is CERTAIN the term appears, $p_t = 1$ and $p'_t = 0$.

We can compute KL divergence for term t:

$$\begin{aligned} KL(P||Q) &= p_t \log \frac{p_t}{q_t} + p'_t \log \frac{p'_t}{q'_t} && (10) \\ &= 1 \times \log \frac{1}{\frac{n_t}{N}} + 0 \times \log \frac{p_t}{q_t} && (11) \\ &= \log \frac{N}{n_t} && (12) \end{aligned}$$

which is exactly the IDF weight of term t.

Compare to alternatives in Amati and Van Rijsbergen (2002) and Ke (2017).

Shannon Entropy

Information / Entropy

Imagine taking a quiz consisting of 3 true/false questions:

- there are 8 (i.e. $2 \times 2 \times 2 = 2^3$) possible sets of answers to the entire quiz;
- Without the ultimate knowledge about what is correct, each set of answers has a likelihood of $p = 1/8$ to be the correct one.

T	T	T	T	F	F	F	F
T	T	F	F	T	T	F	F
T	F	T	F	T	F	T	F
1	2	3	4	5	6	7	8

Table 1: All possible sets of answers to a quiz of 3 binary questions. Each answer set has a probability of 1/8 to be the correct one.

Input Variables

Assume the 3 questions are independent (without an overlap of knowledge), one needs 3 piece of information to answer them.

Therefore, the amount of information required to find the correct answer is proportional to:

$$
\begin{aligned}
H &= 3 \\
&= \log_2 2^3 \\
&= -\log_2 \frac{1}{8} \\
&= -\log_2 p
\end{aligned}
$$

Information / Entropy

Of the overall $-\log_2 \frac{1}{8} = 3$ (bits), each answer set i of the 8, if treated equally likely, requires the following amount to be confirmed or eliminated as the ultimate outcome:

$$H_i = -\frac{1}{8}\log_2 \frac{1}{8} \tag{13}$$

where $\frac{1}{8}$ can be regarded as the probability of i^{th} outcome p_i:

$$H_i = -p_i \log_2 p_i \tag{14}$$

Information / Entropy

Given a set of m mutually exclusive events, its Shannon entropy can be computed by:

$$H = \sum_{i=1}^{m} -p_i \log p_i \tag{15}$$

Back to example of 3 quizzes:

$$\begin{aligned} H &= \sum_{i=1}^{8} -\frac{1}{8} \log_2 \frac{1}{8} && (16) \\ &= 3 && (17) \end{aligned}$$

References

Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, October 2002. ISSN 1046-8188. doi: 10.1145/582415.582416. URL http://doi.acm.org/10.1145/582415.582416.

Weimao Ke. Text retrieval based on least information measurement. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '17, pages 125–132, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4490-6. doi: 10.1145/3121050.3121075. URL http://doi.acm.org/10.1145/3121050.3121075.

References ii

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4): 333–389, April 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL http://dx.doi.org/10.1561/1500000019.

Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986. ISBN 0070544840.