# 相関係数と情報検索のための実証的重みの分析 *

金谷 敦志 † 梅村 恭司‡
豊橋技術科学大学 情報工学系

## 1 はじめに

情報検索において，関連の判断により，単語の重みを評価する手法が Empirical Term Weighting and Expansion Frequency[2] で提案されている．この手法では，単語の重みを参照するために用いられる関数 $\lambda$ のうち，訓練集合から経験的に重み $\hat{\lambda}$ を計算し，情報検索に有用と思われる単語へ直接的に重みをかけるというものである．結果としての重みは，通常 0 から $idf$ の間にあり，$\hat{\lambda} \approx a(tf) + b(tf) \cdot idf$ の関係が得られた．ここで $a(tf)$,$b(tf)$ とは $tf = 0, 1, 2, 3, \geq 4$ それぞれの $tf$ に対する係数である．

本論文では，[2] を基礎として，相関のある単語間について相互情報量を用いた実証的重みの分析を試みた．以下は記号の定義を示す．

- $t$ : 単語
- $d$ : ドキュメント
- $N$ : コレクション内のドキュメントの数
- $N_{rel}, N_{\overline{rel}}$ : 関連のある／ないドキュメントの数
- $tf(t,d)$ : ドキュメント $d$ に含まれる単語 $t$ の数
- $df(t)$ : $t$ を含むドキュメントの数
- $idf(t) \equiv -\log_2 \frac{df(t)}{N}$ : inverse document frequency
- $df(t|rel)$, $df(t|\overline{rel})$: $t$ を含み関連のある／ないドキュメントの数

*Analysis of Empirical Term Weighting for Correlation Coefficient and Information Retrieval

†Atsushi Kanaya (kanaya@ss.ics.tut.ac.jp), Department of Information and Computer Sciences, Toyohashi University of Technology

‡Kyoji Umemura (umemura@tutics.tut.ac.jp), Department of Information and Computer Sciences, Toyohashi University of Technology

## 2 相互情報量と Empirical Term Weighting との関係

[1] において，語 $x$ と $y$ の生起する確率がそれぞれ $P(x)$,$P(y)$ とし，$x$,$y$ が同時に生起する確率を $P(x,y)$ とするとき，2 語の持つ相互情報量 $I(x,y)$ は次式のように定義される．

$$I(x,y) \equiv \log_2 \frac{P(x,y)}{P(x)P(y)}$$

ここで，$x$ の生起する確率は，全ドキュメント数 $N$ と $x$ を含むドキュメント数 $df(x)$ の比と推測し，$P(x) \approx df(x)/N$ と推定する．相互情報量は $x$,$y$ が同時に生起する確率と $x$,$y$ が独立して生起する確率を比較する．ここで，$x = y$ としたときの相互情報量は，

$$\begin{aligned} I(x,x) &= \log_2 \frac{1}{P(x)} \\ &= idf(x) \end{aligned}$$

となり，$\hat{\lambda}$ と同じく $idf$ の一次関数となることから，Empirical Term Weighting の発展としての 2 単語間の線形の関係が存在すると予測される．

## 3 訓練集合を用いた計算

今回の目的としては，$q_i \in$ クエリー中の文字 n-gram，$y \in$ ドキュメント全ての文字 n-gram としたときの $I(q_i, y)$ の高い $y$ に対して $\lambda(y)$ を計算し，$\lambda(y)$ が $I(q_i, y)$ の一次式になるかを確かめることである．対象データには，NTCIR1 の日本語論文アブストラクト 1 万件 (6.5MB) と，対応するクエリー，正解ファイルを用いた．

### 3.1 $df(q_i)$ の計算

$q_i \in$ クエリーの n-gram を取得し，$q_i$ が含まれるドキュメントの数 $df(q_i)$ を計算する．このとき，クエリーに関連のある／ないドキュメント番号を控えておき，後の $df(y|rel), df(y|\overline{rel})$ の計算に用いる．

### 3.2 $df(y)$ の計算

$y \in$ クエリーに関連のある／ないドキュメントのアブストラクトの n-gram を取得し，$y$ が含まれるドキュメントの数 $df(y)$ を計算する．

### 3.3 $I(q_i, y)$ の計算

$I(q_i, y)$ を計算するためには，$q_i$,$y$ を共に含むドキュメントの数 $df(q_i, y)$ を計算する必要がある．相互情報量 $I(q_i, y)$ の計算は，$df(q_i), df(y), df(q_i, y)$ を用いる：

$$I(q_i, y) \approx \log_2 \frac{N \cdot df(q_i, y)}{df(q_i) df(y)}$$

### 3.4 $\lambda(y)$ の計算

相互情報量の小さな単語組合せは膨大にあるため無視し，$I(q_i, y) > 1.0$ となる単語 $y$ に対してのみ，$\lambda(y)$ を計算する．$\lambda(y)$ は likelihood ratio の対数として解釈される：

$$\lambda(y) = \log_2 \frac{P(y|rel)}{P(y|\overline{rel})}$$

分子，分母はそれぞれおおよそ以下のようになる：

$$\begin{aligned} P(y|rel) &\approx \frac{df(y|rel)}{N_{rel}} \\ P(y|\overline{rel}) &\approx \frac{df(y|\overline{rel})}{N_{\overline{rel}}} \end{aligned}$$

### 3.5 実験結果のプロット

相互情報量 $I(q_i, y)$ を横軸，$\lambda(y)$ を縦軸としてプロットした結果の一例を図 1 に示す．$I(q_i, y)$ の増加に伴い，$\lambda(y)$ が線形増加していることが確認された．この結果より，クエリーの単語と共起する確率の高い単語に対して，単語の重みを直接的にかけられると推測する．そして共起する確率の高い単語を用いた検索質問拡張において，正しい重みをかけるという応用が期待される．

## 4 今後の展望

共起する確率の高い単語に対しての重みを直接的にかけられることができる場合，検索質問拡張としての応用により，検索性能の向上が見込まれる．共起する単語対を調べるには計算時間がかかるという欠点はあるが，クエリーの単語と共起する単語を列挙し，$I(q_i, y)$ に対応

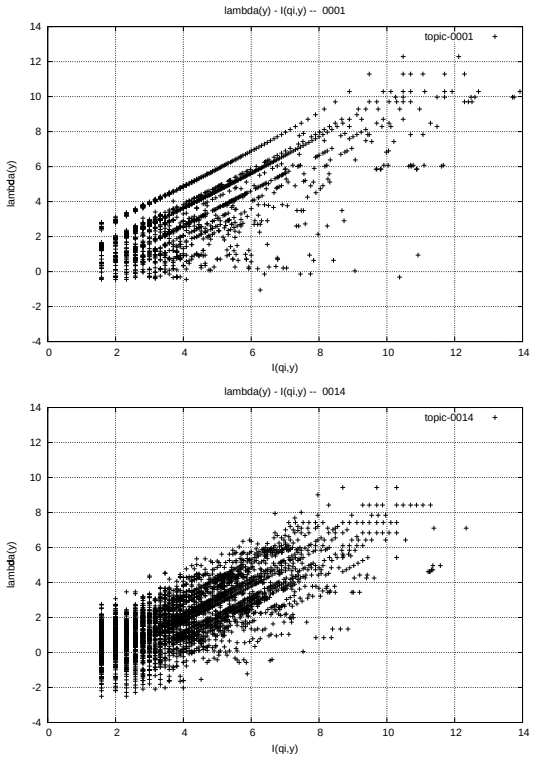


図 1: 相互情報量 $I(q_i, y)$ に対する重み $\lambda(y)$ の分布

する $y$ に対して $\lambda(y)$ を重みをかけて再検索を行い，検索質問拡張の利点である再現率がより向上するかを調べる．

今までの検索質問拡張との違いは，関連性のある単語に対して適切な重みをかけられることが大きな違いであり，今後の予定としては検索性能の向上を実証することである．

## 参考文献

[1] Kenneth W. Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, Vol. 16, No. 1, pp. 22–29, March 1990.

[2] Kyoji Umemura and Kenneth W. Church. Empirical term weighting and expansion frequency. *Workshop SIGDAT, EMLNP2000*, pp. 117–123, 2000.