6M-4

検索質問拡張における実証的重みの検証*

小林　啓一郎†　　金谷　敦志¶　梅村　恭司‡

豊橋技術科学大学 情報工学系†,‡　　住友電工情報システム¶

１．はじめに

情報検索において実証的に重みを計算する手法が,Umemura ら[1]によって提案されている．この手法では，索引語に対して与える重み関数を訓練集合から経験的に求める．金谷ら[2]は，[1]を基礎として，相互情報量を用いた実証的重みの提案をおこない，更にその有効性を示した．しかし、[2]の検証においては，実証的重みの訓練データが 2 万件，検索対象が英語コーパス約 1500 件であり，検証が充分でない．相互情報量を用いた実証的重みの有効性を改めて検証することを目的として，以下の３点について実験を行った．

・訓練データ量による実証的重みへの影響
・大規模コーパスに対する検索
・日本語コーパスに対する検索

２．相互情報量に基づく実証的重み

実証的な重み計算の手法では，訓練データを用いて，文書の適合・不適合の判断を行うことにより索引語の重み関数 $\hat{\lambda}$ を評価する．

相互情報量 $I(x;y)$ は，索引語 x と索引語 y の共起確率と出現確率からなる．

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

索引語 x と索引語 y が完全に独立している場合は，相互情報量は 0 となる．

$$I(x;y) \equiv \log_2 \frac{P(x)P(y)}{P(x)P(y)} = \log_2 1 = 0$$

相互情報量が最大となるのは，2 語が完全に共起する場合であり，相互情報量は以下のように idf となる．

$$I(x;x) \equiv \log_2 \frac{1}{P(x)} = \log_2 \frac{N}{df(x)} = idf(x)$$

このことから，本研究での索引語の重みは，idf を特別な場合として含む相互情報量を用いる．

検索質問中の索引語を $t_q \in q$，文書中の索引語 $t_d \in d$ としたとき，索引語の重み $\hat{\lambda}$ は対数尤度比として推測し，以下のような式で表す．

$$\hat{\lambda}(t_d) = \log_2 \frac{P(t_d \mid \text{適合文書})}{P(t_d \mid \text{不適合文書})}$$

訓練データによって，検索質問中の索引語を $t_q \in q$，検索質問語に対する適合文書中の索引語 $t_d \in d$ の組み合わせについて，相互情報量 $I(t_q;t_d)$ と $\hat{\lambda}$ を求め，I を基にしたビンによるグループ化を経て線形回帰することにより，相互情報量に基づく実証的重みが得られる．この実証的重みを *fit-MI* と呼び，以下の式で表す．

$$\hat{\lambda}(t_q,t_d) \approx a(tf) + b(tf) \cdot I(t_q;t_d)$$

３．実証的重みの計算

NTCIR1 日英論文アブストラクト 33 万件(186MByte)を訓練データとして使用し．この 33 万件の中から 2 万件・10 万件・33 万件を取り出し，それぞれに対して実証的重みの計算を行った．用語の索引付けには，バイグラムを用いた．訓練に使用した質問数は 30 問である．

ビン化の結果を図 1,図 2 に示す．それぞれ 2 万件，33 万件に対する I と $\hat{\lambda}$ の関係である．いずれの結果においても，正の相関があることが見て取れる．これらの結果を線形回帰した結果の傾きと切片は表 1,表 2 のようになった．

これらの重みを用いて検索実験をおこなう．

---

* Verification of Empirical Term Weighting for Query Expansion

† Keiichiro Kobayashi , Department of Information and Computer Sciences, Toyohashi University of Technology

‡ Kyoji Umemura , Department of Information and Computer Sciences, Toyohashi University of Technology

¶ Atsushi Kanaya , Sumitomo Electric Information Systems Co.

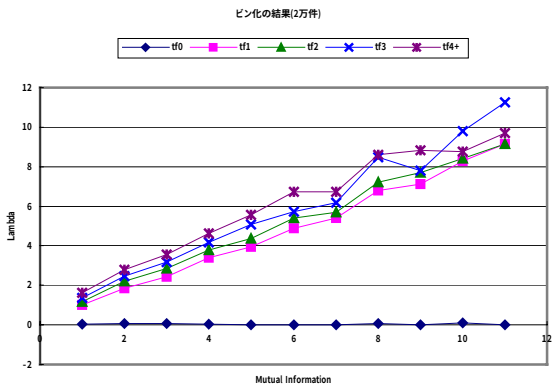

図 1．各 *tf* 値に対する $\hat{\lambda}$ の重み(2 万件)

表 1．各 *tf* 値における初期値 *a* と傾き *b*(2 万件)

tf	*a*(*tf*)	*b*(*tf*)
0	0.061	-0.006
1	0.166	0.789
2	0.432	0.811
3	0.441	0.919
4+	0.785	0.942

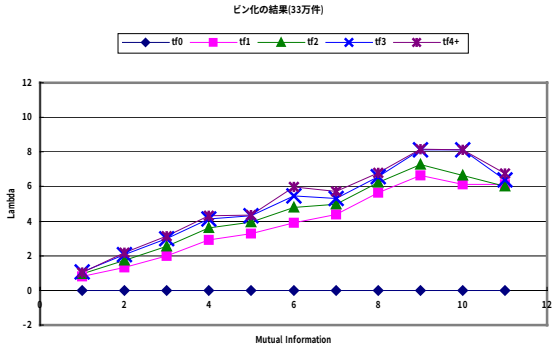

図 2．各 *tf* 値に対する $\hat{\lambda}$ の重み(33 万件)

表 2．各 *tf* 値における初期値 *a* と傾き *b*(33 万件)

tf	*a*(*tf*)	*b*(*tf*)
0	0.011	-0.001
1	0.083	0.657
2	0.386	0.715
3	0.640	0.744
4+	0.623	0.793

4．検索実験

重みの計算を行ったものと同じコーパスに対して，拡張検索実験をおこなった．コーパスサイズは，重みの際と同様に，サイズを 2 万件・10 万件・33 万件と変化させて検証をおこなった．実験に用いた検索質問は 53 件で，重み計算の際に使用したものとは別のものである．

検索方法は，初期検索として *tf*idf* を実行し，その上位文書を適合文書と仮定したうえで，検索質問に対して相互情報量の高い索引語を追加することで，拡張質問とした．拡張検索の際の重みとしては *fit-MI* の他に，*idf* に基づいた実証的重みである *fit-G* に対しても実験を行った．

検索結果の一例を表 3 に示す．なお，評価の尺度は *11pt Average* である．

表 3: 検索結果

重み	20k	100k	330k
*tf*idf*	0.101	0.080	0.115
Fit-G	0.122	0.111	0.203
Fit-MI:20k	0.125	0.119	0.216
Fit-MI:100k	0.124	0.117	0.211
Fit-MI:330k	0.129	0.120	0.215

この実験においては，*fit-G* に比べて良好な結果が得られたことから，日本語・大規模コーパスにおいても，*fit-MI* が有効であることが示唆される．また，実証的重みの計算元となるコーパスサイズは，コーパス全体の一部から求めた重みであっても問題無いと言える．このことから，[2]において検証された *fit-MI* の有効性に関しても妥当であると言える．

5．まとめ

実証的重みについて，大規模コーパスに対する実験をおこない，重みとしての *fit-MI* の有効性を検証した．今後の課題としては，検索質問拡張の際の用語選定方法があげられる．

参考文献

[1] Kyoji Umemura and Kenneth W. Church. Empirical term weighting and expansion frequency. Proceedings of the WVLC-2000, ACL SIGDAT conference EMNLP. pp.117-123.

[2] 金谷敦志，梅村恭司．相関係数を用いた実証的重みの分析と検索質問拡張． 情報学基礎研究会， 第 73 回， pp.17-24， 2003.