LD-001

# 交差確認法に基づく適合性フィードバック
## Relevance feedback with cross validation

藤野 昭典 † 上田 修功 † 斉藤 和巳 †
Akinori Fujino Naonori Ueda Kazumi Saito

## 1. まえがき

確率モデルに基づくテキスト情報検索では，文書の検索順位を決定する尺度を得るのに，文書集合に含まれる各単語がユーザの検索要求に対してどの程度適合するかを推定する必要がある．この単語の適合性を推定する一手法として，適合性フィードバックがある．適合性フィードバックでは，ユーザによって適合と判定された文書集合での単語の出現頻度や単語の出現文書数などの統計情報をもとに単語の適合性が推定される [1, 3, 4]．しかし，ユーザによってフィードバックされる適合文書の数は一般的に少なく，これらの適合文書の統計情報が必ずしも検索対象の適合文書の統計情報を近似しているとは限らない．効果的な検索を実現するには，少ない適合文書から汎化性の高い単語の適合性を推定することが課題となる．この課題に対して本研究では，フィードバックされた適合文書から leave-one-out 交差確認法 (cross validation) により汎化性を考慮した単語の適合性を推定するアプローチを検討した．本稿では，本アプローチによる検索順位尺度の単語の重み推定の最適化と単語選択法，ならびにその効果を検証した結果を述べる．

## 2. 検索モデルと単語の重み推定の最適化

確率モデルに基づくテキスト情報検索では，確率比 (Probability Ratio)[2, 5] に基づいて検索順位が決定される．本研究では，適合クラス $r$ と文書集合全体 $G$ での単語 $i$ の出現確率 $\theta_{ri}, \theta_{Gi}$ の対数確率比を単語の重みとし，以下の式で定義される文書 $n$ の確率比 $PR(n)$ を検索順位尺度 (以下，順位尺度と略記) として用いる．

$$PR(n) = \frac{1}{Z(n)} \sum_{i=1}^{V} tf(n,i) \log \frac{\theta_{ri}}{\theta_{Gi}} \qquad (1)$$

式 (1) 中の $tf(n,i)$ は文書 $n$ での単語 $i$ の出現頻度，$V$ は文書集合群での単語の種類の総数を表す．$Z(n)$ は，文書長により $PR(n)$ の取り得る範囲が変動するのを正規化する項であり，$Z(n) = \sqrt{\sum_{i=1}^{V} tf(n,i)^2}$ とする．適合クラスの単語出現確率 $\theta_{ri}$ は，フィードバックされた適合文書 $n$ と検索質問 $q$ に含まれる単語の出現確率 $P(i|n)$，および平滑化パラメータ $\xi_i$ を用いて Naive Bayes モデルに基づく以下の式で推定する．検索質問に含まれる単語が多く出現する文書は適合クラスに属する確率が高いと考えられるため，検索質問を含めて $\theta_{ri}$ を推定する．

$$\theta_{ri} = \Big\{ \sum_{n \in r,q} P(i|n) + \xi_i \Big\} \Big/ \sum_{i=1}^{V} \Big\{ \sum_{n \in r,q} P(i|n) + \xi_i \Big\} \qquad (2)$$

平滑化パラメータ $\xi_i$ は，順位尺度の汎化性を向上することを目的として，leave-one-out 交差確認法に基づいてフィードバックされた適合文書の確率比 $PR(n)$ の総和を最大にするように決定する．具体的には，$\xi_i = \lambda\theta_{Gi}$ とし，ある適合文書 $n$ の確率比 $PR(n)$ を $n$ 以外の適合文書で推定した $\theta_{ri}^{(-n)}$ をもとに式 (1) から求める場合に，$PR(n)$ の総和が最大になるように $\lambda$ を最適化することで決定する．この推定により，フィードバックされた適合文書の単語出現確率の類似性が高い場合，その確率をより反映した順位尺度が生成され，フィードバックされた適合文書と同じ単語を多く含む文書が上位に検索されるようになる．適合文書の単語出現頻度の類似性は，フィードバック文書集合と検索対象文書集合で相関があると考えられる．このため，単語出現頻度や単語の出現文書数に加えて，平滑化パラメータの推定をもとに適合文書の類似性を考慮することで，検索要求を満たす文書の統計情報により則した検索が期待できる．

## 3. 単語選択による検索

前節では，適合文書の類似性に基づいて単語の重みを一律に調節する手法について述べた．しかし，フィードバックされた適合文書内の記述全てがユーザの検索要求を満たしているとは限らず，文書から適合性の高い単語を選択することで検索性能が高まる可能性がある．そこで，フィードバックされた文書から適合性の高い単語を選択して検索に用いる手法を検討する．まず，単語を選択して検索を行うための順位尺度は，式 (1) に単語の選択を表すパラメータ $\alpha_i \in \{0,1\}$ を導入して定義する．

$$PR'(n) = \frac{1}{Z(n)} \sum_{i=1}^{V} \alpha_i tf(n,i) \log \frac{\theta_{ri}}{\theta_{Gi}} \qquad (3)$$

$\alpha_i$ は，leave-one-out 交差確認法をもとに，適合文書の確率比を大きくするのに寄与する単語を求めることで決定する．すなわち，1 つの文書 $n$ を除いて適合クラス $r$ の単語出現確率 $\theta_{ri}^{(-n)}$ を推定するとき，その出現確率を用いた順位尺度で見積られる文書 $n$ の確率比を大きくする単語 $i$ を求める．フィードバックされたすべての適合文書の確率の総和を大きくする単語 $i$ は，以下の式で定義する単語の寄与度 $\beta_i$ に基づいて選択する．

$$\beta_i = \sum_{n \in r} \frac{1}{Z(n)} tf(n,i) \log \frac{\theta_{ri}^{(-n)}}{\theta_{Gi}} \qquad (4)$$

式 (4) より，$\beta_i$ は単語ごとに適合文書における対数確率比を加算したものであり，$\beta_i$ が大きな単語ほどそれぞれの適合文書の確率比を大きくするのに寄与する．交差法では，単語の選択数が $N_t$ のとき，順位尺度 $PR'(n)$ のパラメータ $\alpha_i$ を寄与度 $\beta_i$ が大きな $N_t$ 個の単語で 1，それ以外の単語で 0 として決定する．$\theta_{ri}^{(-n)}$ を推定する際の平滑化パラメータ $\xi_i$ は，単語によらない一定値 $\xi$ と

†日本電信電話株式会社 NTT コミュニケーション科学基礎研究所

表 1: テストコレクション

| | MED | CRAN | CISI |
|---|---|---|---|
| 文書数 | 1033 | 1400 | 1460 |
| 単語数 | 5462 | 4000 | 5041 |
| 検索質問数 | 30 | 219 | 74 |
| 平均適合文書数 | 23.2 | 8.3 | 42.1 |

し，各フィードバックにおいて文書の確率比の総和が最大になるように決定する．

## 4. 評価実験と考察

### 4.1 実験方法

単語の重みの推定法と単語の選択法の評価実験には，論文の抄録データと検索質問 (query) からなる MED, CRAN, CISI の 3 つのテストコレクションを用いた．評価実験では，各コレクションの抄録を単語の出現頻度ベクトルで表して検索に用いた．単語の出現頻度ベクトルを求める際，冠詞など文書を特徴付ける効果をもたない停止語 (stop word) と，各コレクションにおいて 1 つの抄録のみに出現する低頻度語を取り除いた．各コレクションに含まれる文書数，単語数，検索質問数，ならびに検索質問に適合する平均文書数を表 1 に示す．

適合フードバックに用いる文書は，検索質問と文書の単語出現頻度ベクトルにコレクション全体での $idf(i)$(inverse document frequency) で重み付けを行い，そのコサイン類似度を求める IDF 法で決定した．コサイン類似度が大きな上位 $N$ 個の文書に対するユーザの適合・不適合の判定結果をもとに順位尺度を生成した．ユーザの判定は，コレクションに含まれる適合情報通りに行われるとした．フィードバックされる文書数 $N_{fb}$ は 10, 20, 30 の 3 通りとした．実験には，2 つ以上の適合文書がフィードバックされ，かつ残りの検索対象文書に適合文書が含まれている検索質問を用いた．性能評価は，フィードバック後の残りの検索対象文書に対する 11 点平均適合率 [5] を，各コレクションの検索質問で平均した値を用いた．11 点平均適合率は，0 から 1 までの 0.1 刻みの 11 の再現率レベルにおける適合率を加算平均したものである．平均適合率が大きいほど，適合文書を上位に検索する能力があることを示す．

### 4.2 単語の重み推定の最適化の効果

表 2 は，適合性フィードバックを行わない IDF 法と，$\lambda = 1$ として式 (1) から得られる順位尺度を用いる方法 (FB1)，交差確認法に基づいて最適化した $\lambda$ から得られる順位尺度を用いる方法 (FB2) で検索順位を決定したときの平均適合率を示す．表 2 より，すべての実験で FB2 で検索性能が高い結果がみられた．これは，フィードバックされた適合文書の単語出現確率の類似性を考慮した順位尺度を生成することによって，より高い検索性能が得られることを示している．

### 4.3 単語選択の効果

$\lambda$ の最適化に加えて，式 (3), (4) を用いて単語を選択して順位尺度を生成した場合の検索結果を表 3 に示す．単語の選択法自身を評価するため，交差確認法に基づく方法 (交差法) と，単純に対数確率比 $ratio(i) = \log \frac{\theta_{ri}}{\theta_{Gi}}$ が大きい単語を選択する方法 (確率比) とで，同数の単語を選択した場合の検索性能を比較した．単語の選択数

表 2: 平均適合率での検索方法の比較

| $N_{fb}$ | 方法 | MED | CRAN | CISI |
|---|---|---|---|---|
| 10 | IDF 法 | 0.428 | 0.175 | 0.188 |
| | FB1 | 0.454 | 0.279 | 0.153 |
| | FB2 | **0.571** | **0.354** | **0.237** |
| 20 | IDF 法 | 0.369 | 0.098 | 0.141 |
| | FB1 | 0.464 | 0.251 | 0.121 |
| | FB2 | **0.588** | **0.342** | **0.198** |
| 30 | IDF 法 | 0.219 | 0.078 | 0.124 |
| | FB1 | 0.412 | 0.231 | 0.117 |
| | FB2 | **0.523** | **0.315** | **0.204** |

表 3: 平均適合率での単語選択法の比較: **太字**は 2 つの方法の比較で適合率比が 10%以上高いことを示す．

| | $N_{fb}$ | 方法 | $\gamma = 0$ | 0.5 | 1 |
|---|---|---|---|---|---|
| MED | 10 | 交差法 | 0.541 | 0.559 | 0.564 |
| | | 確率比 | 0.534 | 0.564 | 0.576 |
| | 20 | 交差法 | 0.543 | 0.563 | 0.573 |
| | | 確率比 | 0.537 | 0.581 | 0.592 |
| | 30 | 交差法 | 0.510 | 0.520 | 0.517 |
| | | 確率比 | 0.479 | 0.523 | 0.529 |
| CRAN | 10 | 交差法 | **0.338** | 0.357 | 0.363 |
| | | 確率比 | 0.295 | 0.353 | 0.361 |
| | 20 | 交差法 | **0.297** | 0.319 | 0.320 |
| | | 確率比 | 0.253 | 0.308 | 0.328 |
| | 30 | 交差法 | **0.293** | 0.316 | 0.310 |
| | | 確率比 | 0.243 | 0.297 | 0.306 |
| CISI | 10 | 交差法 | **0.229** | 0.234 | 0.253 |
| | | 確率比 | 0.177 | 0.222 | 0.249 |
| | 20 | 交差法 | **0.209** | 0.209 | 0.214 |
| | | 確率比 | 0.156 | 0.200 | 0.215 |
| | 30 | 交差法 | **0.211** | 0.209 | 0.214 |
| | | 確率比 | 0.160 | 0.201 | 0.214 |

$N_t$ は，各フィードバックにおいて $\beta_i > 0$ となる単語 $i$ の数 $N_\beta$ と $ratio(i) > 0$ となる単語 $i$ の数 $N_{rt}$ を基準として $N_t = (1-\gamma)N_\beta + \gamma N_{rt}$ ($\gamma = 0, 0.5, 1$) から決定した．$N_\beta < N_{rt}$ の関係があり，$\gamma$ が大きいほど順位尺度に用いる単語数は多くなる．表 3 より，単語の選択数が少ない場合に交差法が確率比による方法より検索性能が高い傾向がみられる．これは，交差法では検索性能の向上に寄与する適合性が高い単語を優先的に選択されたことを示している．一方，MED では単語の選択数が多い場合に交差法でやや検索性能が低くなる傾向があった．これは，交差法では寄与率 $\beta_i$ の小さな単語の順位付けは平滑化パラメータ $\xi$ に強く影響されることが一因として考えられる．平滑化パラメータの最適化についてはさらに検討する必要がある．また，表 2 の FB2 との比較により，とくに CISI の結果において単語選択による検索性能の向上の可能性が示された．検索性能を向上させる最適な単語数の決定法などが今後の検討課題である．

## 参考文献


[1] Harman, D.: Relevance feedback revisitied, *Proc. of 15th Ann Int'l SIGIR '92*, 1–10 (1992).

[2] Lavrenko, V. and Croft, W.B.: Relevance Models in informaiton retrieval, in *Language Modeling for information retrieval* (Kluwer academic publishers, Netherlands, 2003), pp. 11–56.

[3] Ng, K: A maximum likelifood ratio information retrieval model, *Proc. of the Eighth Text Retrieval Conference (TREC-8)*, 483-492 (1999).

[4] Zhai, C. and Lafferty, J.: Model-based feedback in the language modeling approach to information retrieval, *Proc. of Tenth International ACM Conference on Information and Knowledge Management (CIKM'01)* , 403–410 (2001).

[5] 徳永健伸: 情報検索と言語処理 (東京大学出版会, 1999).