2F1-01

適合フィードバックにおける非適合文書からの文書検索

Document Retrieval using Feedback of Non-relevant Documents

村田 博士[*1] 小野田 崇[*1] 山田 誠二[*2]
Hiroshi Murata Takashi Onoda Seiji Yamada

[*1](財) 電力中央研究所
Central Research Institute of Electric Power Industry

[*2]国立情報学研究所
National Institute of Informatics

This paper reports a new document retrieval method using non-relevant documents. From a large data set of documents, we need to find documents that relate to human interesting in as few iterations of human testing or checking as possible. In each iteration a comparatively small batch of documents is evaluated for relating to the human interesting. This method is called *relevance feedback*. The relevance feedback needs a set of relevant and non-relevant documents. However, the initial presented documents which are checked by a user don't always include relevant documents. Accordingly we propose a feedback method using information of non-relevant documents only. We named this method *"irrelevance feedback"*. The irrelevance feedback selects a set of documents based on learning result by One-Class SVM. Results from experiments show that this method is able to retrieve a relevant document from a set of non-relevant documents effectively.

1 はじめに

インターネットの発展により，巨大なデータベースにアクセスすることも容易になり，人々が入手できる情報は急激に増大している．しかし，それに伴い本当に必要な情報を見つけ出すことは非常に困難になってきている．このような状況の中，情報検索に関する研究の重要性が増しており，特に文書検索は，インターネット上の情報の多くが文書であることから，その重要性は高く [Yates 99]，様々な研究が TREC(Text Retrieval Conference) [TREC] や NTCIR ワークショップ [NTCIR] などで発表されている．

文書検索の最も基本的な方法として，文書を高次元空間上のベクトルで表現するベクトル空間モデル (VSM) が広く利用されている [Salton 83]．VSM による文書検索システムは，クエリベクトルと文書ベクトルの類似度をベクトル間の内積などの計算により求め，値の大きい，つまりベクトル間のなす角が小さい文書を検索結果として提示するものである．

この VSM に基づくシステムの検索精度をユーザと対話的に改善する手法として，適合フィードバック (relevance feedback) [Salton 71] が提案されている．この手法は，提示された検索文書が求めるものに適合しているか否かの判定をユーザが行い，その判定結果をフィードバックするものである．フィードバック結果を反映させる方法としては，ユーザによる評価を基にクエリベクトルを修正する手法がよく用いられる．

これに対して，ユーザによる評価を学習し，検索対象文書を適合，非適合の 2 つのクラスに分類する分類学習の適用が考えられる [Okabe 01]．また，この分類学習に，データ集合を 2 クラスに分類する能力の高い Support Vector Machine(SVM) を用いる手法も提案されている [Drucker 01] [Onoda 03]．この方法は，ユーザにより適合文書と非適合文書の両方が与えられた場合，ユーザ評価を利用して，効率的な文書検索を実現できる．しかしながら，一般の適合フィードバックでは，適合文書を得るまでに，多くの文書をユーザが評価しなければならず，ユーザの評価が非適合文書のみだった場合は，分類学習を適用することができない．

連絡先: (財) 電力中央研究所 システム技術研究所，〒 201-8511 東京都狛江市岩戸北 2-11-1，TEL:03-3480-2111，FAX:03-5497-0318，murata@criepi.denken.or.jp

このような背景から本稿では，ユーザが与える非適合文書情報に One-Class SVM を適用し，非適合情報を有効に活用して，検索文書中の適合文書を特定する新たなフィードバック手法を提案する．

2 One-Class SVM

学習データが $\mathbf{x}_1, \ldots, \mathbf{x}_\ell$, $\mathbf{x} \in \mathbf{R}$ と表現されるとすると，One-Class SVM [Scholköpf 99] では，ほとんどの学習データが入る小さな領域内で $+1$ の値をとり，それ以外の領域で -1 をとるような関数 f を，学習データと原点の間のマージンを最大化するように生成する．この考えに基づいて，実際には以下の 2 次計画問題を解くことになる．

$$\begin{aligned} \min \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu\ell}\sum_i \xi_i - \rho \\ \text{subject to} \quad & (\mathbf{w}\cdot\mathbf{x}_i) \geq \rho - \xi_i, \\ & \xi_i \geq 0. \end{aligned} \tag{1}$$

ここで，ν は，$\nu \in (0,1)$ となるパラメータである．

w と ρ がこの問題の解とすると，判別関数

$$f(\mathbf{x}) = \mathrm{sgn}\left((\mathbf{w}\cdot\mathbf{x}) - \rho\right) \tag{2}$$

は，ほとんどの学習データ $\mathbf{x}_i$ に対し正の値をとる．

式 (1) で表現される最適化問題に対し，Lagrangian を導入すると，以下の式を得る．

$$\begin{aligned} L(\mathbf{w}, \vec{\xi}, \rho, \vec{\alpha}, \vec{\beta}) = & \\ & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu\ell}\sum_i \xi_i - \rho \\ & - \sum_i \alpha_i((\mathbf{w}\cdot\mathbf{x}_i) - \rho + \xi_i) - \sum_i \beta_i\xi_i \end{aligned} \tag{3}$$

ここで，α_i, β_i は Lagrange 乗数である．この Lagrangian を α_i, β_i について最大化し，主変数 w, ξ, ρ について最小化すると，次式が成立する．

$$\mathbf{w} = \sum_i \alpha_i \mathbf{x}_i, \tag{4}$$

$$\alpha_i = \frac{1}{\nu\ell} - \beta_i \leq \frac{1}{\nu\ell}, \quad \sum_i \alpha_i = 1. \tag{5}$$

式 (4)において，$\{\mathbf{x}_i : i \in [\ell], \alpha_i > 0\}$ となる $\mathbf{x}_i$ を Support Vectors と呼ぶ．式 (4) で表現される w の展開式を判別関数である式 (2) に代入すると以下のようになる．

$$f(\mathbf{x}) = \mathrm{sgn}\left(\sum_i \alpha_i \mathbf{x}_i \cdot \mathbf{x} - \rho\right). \tag{6}$$

式 (4) と式 (5) を式 (3) で表現される Lagrangian に代入すると，以下の双対問題を得ることができる．

$$\min_{\vec{\alpha}} \quad \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j \tag{7}$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq \frac{1}{\nu\ell}, \quad \sum_i \alpha_i = 1. \tag{8}$$

α_i と β_i が 0 でなければ，つまり，$0 < \alpha \leq 1/(\nu\ell)$ ならば，最適解では式 (1) の不等式制約が等式となる．ゆえに，等式を満たす α_i に対して ρ を以下のように計算できる．

$$\rho = (\mathbf{w} \cdot \mathbf{x}_i) = \sum_j \alpha_j \mathbf{x}_j \cdot \mathbf{x}_i. \tag{9}$$

3 非適合フィードバック文書検索

ここでは，前章で述べた One-Class SVM を用いた非適合文書のみからのフィードバック文書検索について述べる．

一般的な適合フィードバックでは，システムがクエリに対する検索を行った結果に対し，ユーザが適合，非適合の判定を行い，その判定結果をシステムにフィードバックすることにより，さらに適合性の高い文書を検索する．これに対して，本研究で扱うフィードバック検索では，クエリに対する検索結果について，ユーザが非適合の判定しかできなかった場合を想定している。以下，非適合文書の情報のみがフィードバックされ，その非適合情報を利用して検索する文書検索を「非適合フィードバック文書検索」と呼ぶ．

SVM に基づく適合フィードバック文書検索では，ユーザにより適合/非適合判定された文書がどちらも存在することを前提としている．つまり，適合/非適合の両文書があることから 2 クラスの分類問題として考えられることができ，SVM によって適合/非適合を識別する分離超平面を生成することができる．しかし，非適合フィードバック文書検索では，2 クラス分類する SVM を利用することは困難である．

しかしながら，非適合文書の情報は確実にフィードバックされるため，この情報を利用して文書検索の効率化が図れる可能性がある．本稿では，One-Class SVM を用いて，非適合文書情報を利用し，文書検索の効率化を図る方法を提案する．

2 章で述べてように，One-Class SVM は，与えられた 1 クラスの領域を明確化できる．この One-Class SVM をユーザが判定した非適合文書群に適用すれば，ベクトル空間モデルを用いて，多次元ベクトル表現された非適合文書群の領域を明確にできる．したがって，ユーザの判定した非適合文書領域に入らない文書をユーザに提示すれば，その提示文書にはユーザが適合と判定する文書が入る可能性が高くなり，効率的な文書検索が期待できる．

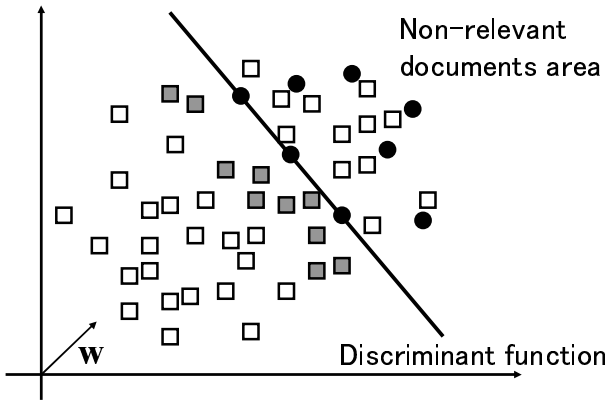

図 1: 非適合フィードバックにより提示される文書 (灰色の□)：●はユーザが非適合と判定した文書．□は未判定の文書．

One-Class SVM による非適合フィードバックは，以下に示す手続きで検索を行う．

Step 1: 初期検索
ベクトル空間モデルを用い，ユーザが要求した質問に対し，検索を行い，文書ベクトルとユーザの質問であるクエリベクトルとのコサイン距離によってその類似度を測り，文書を順位付けする．類似度の高い上位 N 文書をユーザに提示する．

Step 2: ユーザによる判定
Step 1 で提示された文書に対し，ユーザは適合，非適合の判定を行う．ここでユーザの判定が適合/非適合の両方を含む場合，通常の適合フィードバック検索へ移行する．

Step 3: 非適合文書領域の明確化
ユーザが判定した非適合文書を用い One-Class SVM の学習を行い，判定した非適合文書中から非適合文書を覆う境界面を明確化する．

Step 4: 検索
ユーザが判定していない文書を多次元ベクトル空間上にマッピングし，決定された境界面との距離を計算．非適合文書領域内になく，境界面に近い，上位 N 文書をユーザに提示し (図 1 参照)，**Step 2** へ．

非適合フィードバック検索は，より早くユーザに適合と判定される文書を提示することが目的である．**Step 4** で非適合文書領域内になく，境界面に近い文書を提示するのは，この場合，非適合文書であっても，少なくともユーザの与えたクエリは含んでいるため，非適合文書領域外の境界近傍には，非適合文書ではないが，クエリを含んでいる文書が多数存在すると考えられるためである．仮に，最も非適合文書領域から遠い文書を選ぶと，その文書は，ユーザの与えたクエリベクトルとは全く無関係な文書となる可能性が高くなる．

非適合フィードバックによって，ユーザが適合と判定できる文書が提示できれば，**Step 2** にあるように，通常の SVM に基づく適合フィードバック文書検索が利用できる．

4 検索実験

4.1 実験条件

3 章で提案した One-Class SVM による非適合フィードバック文書検索手法の有効性を検討するための実験を行った．

実験用データには，文書検索に関する国際会議 TREC [TREC] で広く使用されている英字新聞記事 (The Los Angels Times, 約 13 万記事，平均単語数 526 語) を使用した．このデータには検索要求文とその要求に適合する文書集合が提供されており，本研究でもこれをクエリとして用いている．

文書ベクトルの算出には，文献 [Schapire 98] を参考に一般的な TFIDF [Yates 99] を改良したものを用いた．具体的には以下の計算式を使った．

$$w_t^d = L * t * u$$

$$L = \frac{1 + \log(tf(t,d))}{1 + log(\text{average of } tf(t,d) \text{ in } d)} \quad \text{(TF)}$$

$$t = \log(\frac{N+1}{df(t)}) \quad \text{(IDF)}$$

$$u = \frac{1}{0.8 + 0.2\frac{uniq(d)}{\text{average of } uniq(d)}} \quad \text{(normalization)}$$

- w_t^d：文書 d における単語 t の重み．
- $tf(t,d)$：文書 d における単語 t の出現頻度
- N：データ集合内の文書総数
- $df(t)$：単語 t を含む文書数
- $uniq(d)$：文書 d における単語の異なり数（種類）

3 章の **Step 1** で述べたフィードバックに用いる文書数 N は，10，20 とした．また，One-Class SVM のパラメータ ν は 0.01 とし，与えられた非適合文書をほぼすべて含む境界を発見するようにした．One-Class SVM の学習，領域の発見には，LibSVM [LIBSVM] を用いて実験を行った．

提案手法の有効性を示すため，フィードバックを行わないベクトル空間モデルに基づく文書ベクトルとクエリベクトルとのコサイン距離での順位付けを基本手法として採用した．また，フィードバック手法との比較を行うため，適合文書情報が無くても動く Rocchio-based フィードバック手法 [Salton 71] を比較のためのフィードバック手法として用いた．

Rocchio-based フィードバック手法は，クエリベクトル (Q_i) を下式により更新し，検索精度を向上させる手法である．

$$Q_{i+1} = Q_i + \alpha \sum_{x \in R_r} x - \beta \sum_{x \in R_n} x, \qquad (10)$$

ここで，R_r は i 回目において検索され，適合と判定された文書集合，R_n は i 回目において検索され，非適合と判定された文書集合である．また，α，β は定数であり，それぞれ関連文書，関連のない文書をどの程度重要視するかを調整する．本稿では，経験的によいとされる $\alpha = 1.0$，$\beta = 0.5$ を採用して実験を行った．

表 1: 繰り返し回数に対する適合文書数 (提示文書数 20)

topic 306

繰り返し回数	One-Class	VSM	Rocchio
1	1	1	0
2	-	-	0
3	-	-	0
4	-	-	0
5	-	-	0

topic 343

繰り返し回数	One-Class	VSM	Rocchio
1	1	0	0
2	-	0	0
3	-	0	0
4	-	1	0
5	-	-	0

topic 383

繰り返し回数	One-Class	VSM	Rocchio
1	1	0	0
2	-	1	0
3	-	-	0
4	-	-	0
5	-	-	0

表 2: 繰り返し回数に対する適合文書数 (提示文書数 10)

topic 306

繰り返し回数	One-Class	VSM	Rocchio
1	1	0	0
2	-	0	0
3	-	1	0
4	-	-	0
5	-	-	0

topic 343

繰り返し回数	One-Class	VSM	Rocchio
1	0	0	0
2	1	0	0
3	-	0	0
4	-	0	0
5	-	0	0

topic 383

繰り返し回数	One-Class	VSM	Rocchio
1	0	0	0
2	1	0	0
3	-	0	0
4	-	1	0
5	-	-	0

4.2 実験結果

本研究での検索実験では，初期クエリベクトルによる検索結果の少なくとも上位 30 文書には，適合文書が含まれない 3 つの topic を用いた．実験方法は，ユーザが N 文書ずつ適合／非適合の判定を行う場合に，何回目のフィードバックで適合文書を含む N 文書の群を提示できるかを評価した．

提案手法 (One-Class)，初期クエリベクトルによる検索方法 (VSM)，および Rocchio-based フィードバック手法についての繰り返し回数に対する適合文書数を，提示文書数が 20 文書の場合を表 1 に，提示文書数が 10 文書の場合を表 2 に示す．

表 1 中の繰り返し回数 1 とは，ユーザが初期検索結果 20 文書を判定した後を表す．つまり，繰り返し回数 1 の時点では，ユーザは全部で 40 文書を見ていることとなり，繰り返し回数 2 の時点では，60 文書を見ていることとなる．表 2 の場合も同様に，繰り返し回数 1 では，ユーザは初期検索結果 10 文書を判定した後であり，全部で 20 文書を見ていることとなる．

表 1 より，提案手法を用いた場合，ユーザは 40 文書見た時点で，適合文書を見つけることができていることがわかる．つまり，初期検索結果の 20 文書が非適合であるとの判定をすれば，次の検索結果の中で，ユーザは適合文書を見つけることができる．初期クエリベクトルとの類似度から検索した場合には，topic 306 が 40 文書を見た時点で適合文書を見つけられるものの，topic 343 では 100 文書，topic 383 では 60 文書を見た時点でないと適合文書を見つけることができない．一方，従来フィードバック手法である Rocchio-based フィードバック手法を用いた場合は，topic 306，343，383 のどれについても，ユーザがシステムの提示する 120 文書を見ても，適合文書を見つけることができないことがわかる．これは，式 (10) で表現されるフィードバック手法に起因すると考えられる．式 (10) では，適合文書があった場合，その適合文書に含まれるタームが強調され，有効なクエリベクトルが生成できる．しかし，本研究のように，非適合文書のみが与えられる場合，式 (10) の右辺第 3 項のマイナス項のみが変化し，コサイン距離がどの文書も殆ど変わらないような状況になってしまう傾向があると考えられる．そのため，表 1 に示した結果となったと考えられる．

ユーザへの提示文書数を 10 文書とした場合の結果を表した表 2 を見ると，topic306 については，提案手法を用いることで，10 文書を判定するだけで適合文書を見つけることができている．これは，本手法において，早い段階で非適合文書の情報をフィードバックすることの有効性を示している．

5 おわりに

本稿では，文書検索の検索効率向上のための One-Class SVM に基づく非適合フィードバック手法を提案した．

TREC の新聞記事のデータを用いた実験において，提案手法の有効性を検証したところ，初期検索結果の少なくとも上位 30 文書中には適合文書が含まれないという条件の下では，本稿で提案した非適合フィードバック手法が，クエリベクトルと文書ベクトルとのコサイン距離での検索方法や Rocchio-based 適合フィードバック検索手法より，検索効率が高いことを示した．

通常，大量のデータ中よりユーザの要求する情報を効率的に見つける問題では，初期にユーザから得られる情報は，ユーザの欲している情報では無いという否定的な情報のみであることが多い．今後は，実際のさまざまな問題において，否定的な情報を効率的に利用する方法について検討していく．

参考文献

[Drucker 01] Drucker, H. Shahrary, B. and Gibbon, D. C.: Relevance Feedback using Support Vector Machines, in *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 122–129, (2001).

[LIBSVM] http://www.csie.ntu.edu.tw/˜cjlin/libsvm/.

[NTCIR] http://research.nii.ac.jp/ntcir/index-ja.html.

[Okabe 01] Okabe, M. and Yamada, S.: Interactive Document Retrieval with Relational Learning, in *Proceedings of the 16th ACM Symposium on Applied Computing*, pp. 27–31, (2001).

[Onoda 03] Onoda, T. Murata, H. and Yamada, S.: Relevance feedback with active learning for document retrieval. In *International Joint Conference on Neural Networks 2003*, pp. 1757–1762, (2003).

[Salton 71] Salton, G. ed.: *Relevance Feedback in Information Retrieval*, pp. 313–323. Englewood Cliffs, N.J.: Prentice Hall, (1971).

[Salton 83] Salton, G. and McGill, J.: *Introduction to Modern Information Retrieval*. McGraw-Hill, (1983).

[Schapire 98] Schapire, R. Singer, Y. and Singhal, A.: Boosting and Rocchio Applied to Text Filtering, in *Proceedings of the Twenty-First Annual International ACM SIGIR*, pp. 215–223, (1998).

[Scholköpf 99] Schölkopf, B. Platt, J. Shawe-Taylor, J. Smola, A. and Williamson, R.: Estimating the Support of a High-dimensional Distribution, TR 87, Microsoft Research, (1999).

[TREC] http://trec.nist.gov/.

[Yates 99] Yates, R. B. and Neto, B. R.: *Modern Information Retrieval*. Addison Wesley, (1999).