3D1-04

SVMに基づく対話的文書検索におけるカーネルの提案

A Proposal of a Kernel suitable for Interactive Document Retrieval Based on SVMs

村田 博士[*1*3] 小野田 崇[*1] 山田 誠二[*2*3]
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 Infomatics

[*3]総合研究大学院大学
The Graduate University for Advanced Studies (SOKENDAI)

This paper describes an application of SVMs (Support Vector Machines) to interactive document retrieval using active learning. We show that SVM-based retrieval have an association with conventional relevance feedback by comparative analysis. We propose a cosine kernel which has the meaning equal with cosine similarity suitable for SVM-based interactive document retrieval from the analysis. We confirm the effectiveness of this method and experimentally compared it with conventional system in several representations of document vectors.

1. はじめに

文書検索における検索精度をユーザと対話的に改善する手法として，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，適合性フィードバック (relevance feedback) [1] が提案されている．我々は，この適合フィードバックを対話的分類学習として捉え，現在最も性能の高い分類学習アルゴリズムの一つであるサポートベクターマシン：SVM (Support Vector Machines) を適用し，ユーザに判定してもらう文書を能動的に選択する能動的文書提示を実現している [2]．本研究では，SVMにおける距離を用いた適合度を定式化し，適合フィードバックで一般的に用いられるRocchioの手法 [1] との比較分析を行う．そこから得られた知見より，文書検索に適したカーネルを提案し，文書検索用のデータセットを用いて検索性能改善の効果について明らかにする．

2. SVMに基づく適合フィードバック文書検索

図1にSVMに基づく適合フィードバックの概念図を示す．図中の○と●は，それぞれ判定済みの適合文書と非適合文書であり，判定済み文書ベクトルを $\mathbf{x}_i$，クラスラベルをそれぞれ $y_i = 1$，$y_i = -1$ とする．このとき図1中の◎で表される未判定文書ベクトル $\mathbf{x}$ の判別超平面からの符号付距離は，

$$\frac{\mathbf{w}\cdot\mathbf{x}+b}{||\mathbf{w}||} = ||\mathbf{x}||\cos\theta_w + \frac{b}{||\mathbf{w}||} \tag{1}$$

となる．ここで，θ_w は $\mathbf{w}$ と $\mathbf{x}$ のなす角である．

一方，SVMにおけるマージン最大化を，$\alpha_i \geq 0$ のLagrange乗数を用いた最大化問題で表したとき，次式が成立する．

$$\mathbf{w} = \sum_{i=1}^{\ell} \alpha_i y_i \mathbf{x}_i \tag{2}$$

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

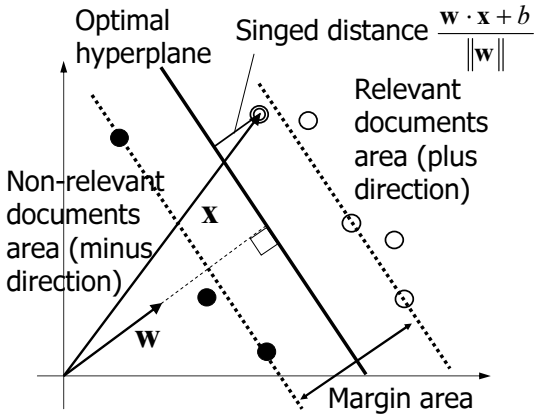

図 1: SVMに基づく適合フィードバック

$\mathbf{w}$ は，適合文書のラベルが $y_i = 1$，非適合文書のラベルが $y_i = -1$ なので，

$$\mathbf{w} = \sum_j \alpha_j \mathbf{x}_j - \sum_k \alpha_k \mathbf{x}_k \tag{3}$$

となる．ここで，j は適合文書，k は非適合文書を示す添字である．

これに対し，Roochioの方法でのクエリベクトル更新式は，次のようになる．

$$Q_{m+1} = Q_0 + \sum_j \beta \mathbf{x}_j - \sum_k \gamma \mathbf{x}_k \tag{4}$$

ここで，Q_0 は初期クエリベクトル（ユーザが最初に与えたクエリのベクトル）である．

式 (3) と式 (4) を比較すると，SVMに基づく適合フィードバックのベクトル $\mathbf{w}$ の式は，Rocchioのクエリベクトル更新式における初期クエリベクトルがゼロベクトルの場合と同等であり，式 (3) の $\mathbf{w}$ をRocchioベース適合フィードバックのクエリベクトルと捉えることができる．

3. 比較分析に基づくカーネルの提案

SVMによる手法での適合度の評価は，$\mathbf{w}$ と bが判定済み文書ベクトルにより一意に決まることから，式 (1) より，$||\mathbf{x}|| \cos\theta_w$ を評価していることになる．一方，Rocchioの手法では文書ベクトルとクエリベクトルとのコサイン類似度，つまり $\cos\theta_w$ を評価している．

以上の比較分析から，SVMによる手法では，対象となる文書ベクトルが大きい，つまり $||\mathbf{x}||$ が大きいと，適合度が高くなることになる．これは，$\mathbf{w}$ との θ_w が小さい未判定文書より，$||\mathbf{x}||$ が極端に大きい未判定文書のほうが適合度が高くなることを意味する．このようなことを避けるために，適合度をベクトル空間モデルで一般的である純粋なコサイン類似度，つまり $\cos\theta_w$ のみにしたほうがよいと考えられる．そのためには，$||\mathbf{x}||$ を定数にする必要がある．これを実現する最も簡単な方法としては，文書を単位ベクトルに正規化すればよい．

一方，SVMで用いられるカーネル $K(\mathbf{x}, \mathbf{x}')$ として，二つのベクトル $\mathbf{x}$ と $\mathbf{x}'$ のなす角を θ としたときのコサイン類似度を用いることを考える．このとき，

$$K(\mathbf{x}, \mathbf{x}') = \cos\theta = \frac{\mathbf{x} \cdot \mathbf{x}'}{||\mathbf{x}||||\mathbf{x}'||} \tag{5}$$

となる．この式を見ると，ベクトルのコサイン類似度をカーネルに用いることは，文書ベクトルの単位ベクトル化と同じであることがわかる．このコサイン類似度を用いたカーネルを，コサインカーネルと呼ぶこととする．コサインカーネルを用いることにより，SVMにおける距離を用いた適合度の評価においても，ベクトル空間モデルと同様の評価を行うことになる．

4. 実験

コサインカーネルの有効性検討のため，文書検索の評価実験で広く使用されている，国際会議 TREC[*1]の第 6 回から第 8 回の ad hoc タスクで使用された約 53 万の新聞記事文書からなるデータセットを用いて実験を行った．SVMの分類性能は文書ベクトル表現のベクトル空間に依存するため，提案するコサインカーネルによる性能の比較評価を行うのと同時に，文書ベクトル表現の違いによる比較評価を行った．各文書ベクトル表現は，Boolean，TF と TFIDF の 3 種類を比較した．TFIDF は，次の計算式を使った．

$$w(t,d) = \frac{\log(\mathrm{tf}(t,d)+1)}{\log(\mathrm{uniq}(d))} \log\frac{N}{\mathrm{df}(t)} \tag{6}$$

ここで，$w(t,d)$：文書 d における単語 t の重み，$\mathrm{tf}(t,d)$：文書 d における単語 t の出現頻度 (TF)，N：データ集合内の文書総数，$\mathrm{df}(t)$：単語 t を含む文書数，$\mathrm{uniq}(d)$：文書 d における単語の異なり数（種類）をそれぞれ表している．

SVM は LibSVM [*2] を用いて実装した．

検索性能を評価する指標としては，各フィードバック回数で終了した場合に，提示された全文書中の累積適合文書の割合 P を用いた．具体的には，ユーザが一回に判定する文書数を S，フィードバック回数を M，最終的に表示する文書 (フィードバック後に学習された SVM において，適合文書領域側にある最適分離超平面からの距離が遠い未判定文書) の数を H とすると，H=S として，最終的にユーザに提示された文書数 $S(M+1)$ とその文書中の適合文書の数 R から，$P = \frac{R}{S(M+1)}$ で計算される．

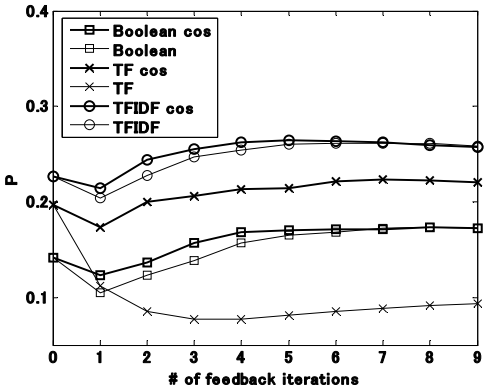

図 2: 提示文書数 $S = 10$ のときの検索性能 P

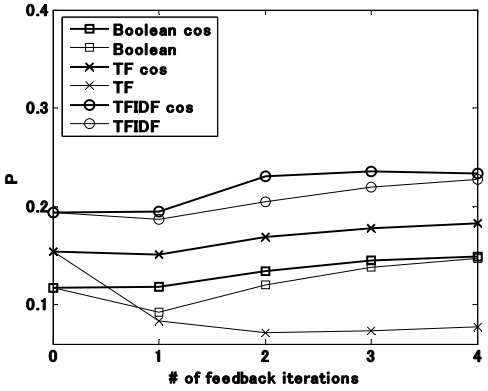

図 3: 提示文書数 $S = 20$ のときの検索性能 P

判定文書の上限は 100 文書に設定した．具体的には，フィードバック時の提示文書数 S を 10，20 とし，それぞれのフィードバック回数 M は，1〜9, 1〜4 へ変化させてパフォーマンスを調べた．また，最終的に提示数する文書数は P の算出でも述べたように $H = S$ とした．

提示文書数 S が 10 と 20 のときの検索性能 P を図 2 と図 3 に示す．ここで，フィードバック回数 0 は，初期検索時の性能を表す．また，太線がコサインカーネルを，細線が線形分離した場合を表す．

図から，提案するコサインカーネルによって，すべての文書表現で性能が向上し，特に TF の性能が大きく向上することがわかる (図中の×のグラフの比較)．

参考文献

[1] G. Salton Ed.: “Relevance feedback in information retrieval”, pp. 313–323, Englewood Cliffs, N.J.: Prentice Hall (1971).

[2] T. Onoda, H. Murata, and S. Yamada, “SVM-based Intaractive Document Retrieval with Active Learning”, New Generation Computing, **26**, 1, pp. 49–61 (2008).

∗1 http://trec.nist.gov/

∗2 Kernel-Machines: http://www.kernel-machines.org/.