The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

SVMに基づく適合フィードバックにおける探索空間縮小効果の実験的分析

Experimental Analysis of Effect of Reducing Search Space for Relevance Feedback Based on SVM

今村優斗　西垣貴央　小野田崇
Yuto Imamura　Takahiro Nishigaki　Takashi Onoda

青山学院大学 理工学部　経営システム工学科
Department of Industrial and Systems Engineering, College of Science and Engineering, Aoyama Gakuin University

SVM (Support Vector Machines) based relevance feedback has proposed. Vector space model is often used for document retrieval. At that time, the search space may become very large because the document vectors have too many attributes which included in the documents on the database. It is presumed that search performance decreases due to the presence of meaningless attributes. However, it is not clear the effect of the size of the search space for the search performance. Therefore, we compared among eight search spaces and conducted an experiment. The search space based on attributes included in the presentation documents improved search performance.

1. はじめに

企業において、自社の特許明細書に類似する他社の特許明細書の検索を行う場面や、研究者が論文を検索する際には、ユーザは文書検索により、より多くの適合文書 (ユーザが求めている文書) を獲得する必要がある。しかし、ユーザが検索クエリ (検索キーワード) を入力し、それに対して検索システムが検索結果を提示するという 1 回だけの検索により、多くの適合文書を獲得することは容易ではない。そのため、より多くの適合文書を獲得したい検索タスクでは、検索結果をユーザが評価し、その評価をもとに再検索を行う、対話的文書検索を行うことが現実的である [1]。この対話的文書検索の 1 つに、検索結果である文書をユーザに提示し、ユーザが適合文書・非適合文書の判定を行い、その判定結果をもとに再検索を繰り返す適合フィードバックがある。適合フィードバックを対話的分類学習としてとらえ、分類学習アルゴリズムの 1 つであるサポートベクターマシン：SVM(Support Vector Machines) を適用する方法が提案されている [2, 1]。

2. 関連研究

2.1 SVM に基づく適合フィードバック

参考文献 [1, 2] に基づき、SVM に基づく適合フィードバックの概要を説明する。

図 1 に SVM に基づく適合フィードバックの概念図を示す。$\boldsymbol{x}$ はデータベース上の文書ベクトルである。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る。ここで◎は未判定文書である。ことのき学習データより得られる SVM の判別超平面 (識別関数) は次の式により表すことができる。

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b \tag{1}$$

ここで $\boldsymbol{w}$ は判別超平面の法線ベクトル (線形識別器の重みベクトル)、非負値である b はバイアス項と呼ばれる定数であり、$\boldsymbol{x}$ は文書ベクトルを示している。

判別超平面と文書ベクトル $\boldsymbol{x}$ の距離は、

$$\frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b}{||\boldsymbol{w}||} \tag{2}$$

連絡先: 今村優斗，青山学院大学 理工学部
経営システム工学科, a5714015@aoyama.jp

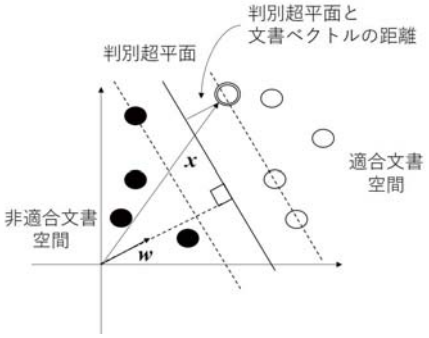

図 1: SVM に基づく適合フィードバックの概念図

となる。SVM に基づく適合フィードバックでは、この判別超平面からの距離を適合度とする。各未判定文書における適合度を計算し、適合文書側にある未判定文書のうち、適合度の大きいものから順に提示をする。

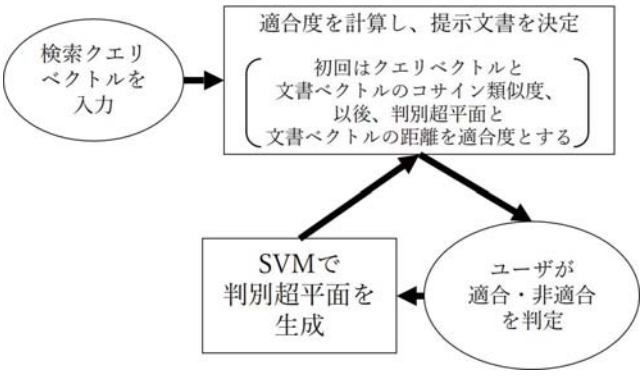

図 2: SVM に基づく適合フィードバックの流れ

図 2 に、SVM に基づく適合フィードバックの流れを示す。初回のフィードバックでは適合度として、文書ベクトルと検索クエリベクトルとのコサイン類似度を使用する。

2.2 SVM に基づく適合フィードバックの課題

文書ベクトルは文書データベース内の文書に含まれる語彙の数だけの属性を持つ。このため、大規模データベースにおける文書の空間ベクトル表現では、数万〜数十万と、膨大な属性をもつベクトルとなってしまい、探索空間が非常に大きくなってしまう。各文書に含まれている属性の数は、データベースに含まれている全文書中で使われている属性の数よりもとても少なく、文書ベクトルは要素に 0 をもったスパースなベクトルとなる。文書全体をこのようなベクトルで表現すると、膨大な記憶容量が必要になると同時に、適合度の計算を行う際の探索空間が大きくなることにより、計算コストも高くなってしまう。また、文書中に含まれる無意味な属性の存在が検索に影響を与え、検索性能を下げてしまうという現象が起こることがある [3]。

これまでの SVM に基づく適合フィードバックの研究では、従来の適合フィードバックの手法との比較実験や、文書ベクトル表現の違いについての考察、文書データベース中の適合文書の割合を変化させた実験などは行われてきているが、これらの研究では SVM に基づく適合フィードバックにおける探索空間縮小効果は考慮されていない [1]。

そこで本論文では、探索空間縮小を縮小することにより、検索性能を改善することができるのかどうかを検証する必要があると考え、複数の探索空間縮小手順を用意し、実験的に分析した。

3. 探索空間縮小手順

本研究では、ランダムに属性を選択する手順 4 つと提示文書の属性のみを用いる手順 3 つ、これに通常の SVM に基づく適合フィードバックを加えた 8 つの手順で実験を行った。

3.1 ランダムな属性選択

ランダムに属性を選択する手順として以下の 4 つの手順を用意した。

- 事前にランダムに 10000 の属性を選択し、その空間に基づき適合フィードバックを行う手順 SVM(random 10000)
- 事前にランダムに 5000 の属性を選択し、その空間に基づき適合フィードバックを行う手順 SVM(random 5000)
- 事前にランダムに 1000 の属性を選択し、その空間に基づき適合フィードバックを行う手順 SVM(random 1000)
- 各フィードバックでランダムに属性を選択し、探索空間を広げていく手順 SVM(random increase)

それぞれ 3 回ずつ実験を行い、その平均値を実験結果に記している。各フィードバックでランダムに属性を選択する手順は、次に述べる提示文書に使用されている属性のみを用いる手順と比較するため、探索空間の大きさが同じになるように各フィードバック回で増やす属性の数を設定した。

3.2 提示文書に使用されている属性のみを用いる手順

次に、システムがユーザに提示した文書で使用されている属性のみを用いて、探索空間を縮小する手順を説明する。以下の 3 つの手順を用意した。

- 提示文書中の適合・非適合文書両方に使用されている属性のみを用いる手順 SVM(reduction)
- 提示文書中の適合文書に使用されている属性のみを用いる手順 SVM(reduction relevant)
- 提示文書中の非適合文書に使用されている属性のみを用いる手順 SVM(reduction non-relevant)

これら手順では、フィードバック $(m-1)$ 回目までにシステムの提示した文書で使用されていた属性のみを用いた探索空間を作成し、その探索空間に基づいて m 回目の検索を行う。例えば 2 回目のフィードバックでは、それまでの 0 回目、1 回目のフィードバックでの提示文書で使用されていた属性のみを用いて、再検索を行う。「システムが適合と判断をしたが、実際は非適合であった文書 (非適合文書)」と、「システムが適合と判断をし、実際に適合であった文書 (適合文書)」に使われていた属性のみを用いることで、適合・非適合の判断に有用であろうと考えられる属性を選択していく。大規模データベースでは、データベース中の全文書数と比べユーザに提示する文書数は非常に少ないため、これら 3 つの手順を使用することで、使用する属性を大きく減らすことができる。したがって、探索空間を大幅に縮小することが可能である。

4. 実験

4.1 実験条件

実験用のデータには参考文献 [1] 等文書検索の評価実験で使用されている、国際会議 TREC の adhoc タスクで使用されたデータセットを用いた。このうち、約 13 万の新聞記事の文書データからなるデータセットを用いた。100 個の検索課題（トピック）が用意されている。それらの検索課題にはそれぞれ適合文書が用意されており、本論文ではこの中から十分に適合文書が用意されている 20 個の検索課題を用いた。検索課題は、複数の単語で表現されている。また、各文書とクエリには、smart system のリスト [*1] を使った stopword の除去と Porter Stemming Algorithm による stemming[*2] を行った。

SVM の分類性能はベクトル空間の文書ベクトル表現に依存するため、TF と TF-IDF の 2 種類で比較を行った。TF-IDF は一般的に使われている次の計算式を使用した [1]。

$$w(t,d) = \frac{\log(tf(t,d)+1)}{\log(uniq(d))} \log \frac{N}{df(t)} \tag{3}$$

- $w(t, d)$：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。
- $df(t)$：属性 t を含む文書数。
- $uniq(d)$：文書 d における属性数

文書ベクトルの属性数は約 17 万である。SVM に基づく適合フィードバックの初回フィードバック次では、コサイン類似度を用いて適合度を計算し、初回の提示文書を選んでいる。その際にはすべての属性を使用している。それ以降のフィードバックでは、判別超平面からの距離が適合文書領域側で最も遠い文書からユーザに提示する。SVM の学習にはすべての提示文書を使用している。

∗1 http://www.lextek.com/manuals/onix/stopwords2.html
∗2 http://tartarus.org/?martin/PorterStemmer/

4.2 評価指標

参考文献 [1] で用いられている指標を基にした。検索性能を評価する指標として、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 1 回に判定する文書数を S, フィードバック回数を m(初回は $m=0$) として、そのフィードバック回までにユーザに提示された文書の数 $S(m+1)$ とその文書中の適合文書の数 R から、

$$\mathrm{P} = \frac{R}{S(m+1)} \tag{4}$$

で計算した。また、学習性能の評価は「上位 30 文書における精度 P30」を適用する。 これにより、最終的に生成された分類器の検索における学習性能が比較できる。以下の式で計算した。p は各回の上位 30 文書中の適合文書数である。

$$\mathrm{P30} = \frac{p}{30} \tag{5}$$

各フィードバック回の提示文書数 $S=10$ 、最大フィードバック回数 $M=9$ とした。各検索課題で、M 回再検索を行い各回で S 個の文書をユーザに提示する。各フィードバック回で評価指標を計算し、検索課題全ての平均値を計算し、グラフにした。

4.3 実験結果

表 1 は、提示文書に使用されている属性のみを用いて探索空間を縮小する手順で、SVM を使用した適合フィードバックを行った際の属性数の推移の平均である。適合・非適合文書に使用されている属性のみを用いた場合と、適合文書のみ、非適合文書のみの属性を用いた場合の計 3 つの手順の属性数の推移を示している。各条件での全検索課題の各フィードバック回における探索空間の大きさの平均をとっている。縮小前の属性数は 172896 である。元の探索空間と比べ、非常に小さい探索空間で検索が行われていることがわかる。なお、各フィードバック回でランダムに属性を選択し、属性数を増やしていく手順では、このうち適合・非適合文書の属性を使用した場合の属性数の平均と同様の探索空間の大きさで適合フィードバックを行うように設定し、実験をおこなった。

表 1: 属性数の推移

提示文書に使用されている属性を用いる手順

適合・非適合文書	$m=1$	$m=2$	…	$m=8$	$m=9$
TF	1154	2103	…	5340	5662
TF-IDF	447	1316	…	4249	4581
適合文書のみ	$m=1$	$m=2$	…	$m=8$	$m=9$
TF	581	798	…	1928	2016
TF-IDF	223	601	…	1803	1884
非適合文書のみ	$m=1$	$m=2$	…	$m=8$	$m=9$
TF	676	1425	…	4241	4656
TF-IDF	245	653	…	2672	2956

図 3、図 4、図 5、図 6 の凡例において、SVM と表記されているグラフは探索空間を縮小していない通常の SVM に基づく適合フィードバックのグラフである。SVM(reduction) は提示文書全てで使用されていた属性のみを用いる手順のグラフである。SVM(reduction relevant) は提示文書のうち、適合文書で使用されている属性のみを用いる手順のグラフであ

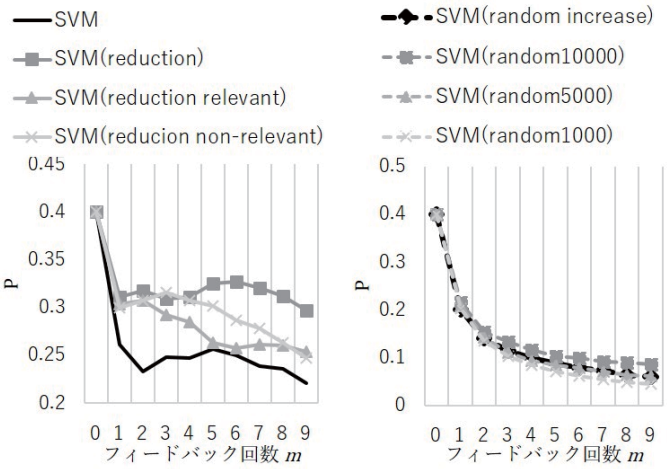

図 3: TF における検索性能 P

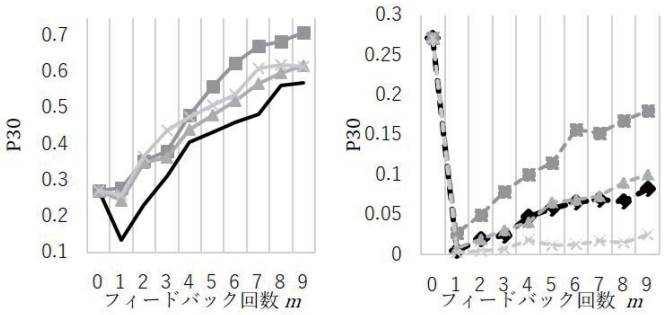

図 4: TF における学習性能 P30

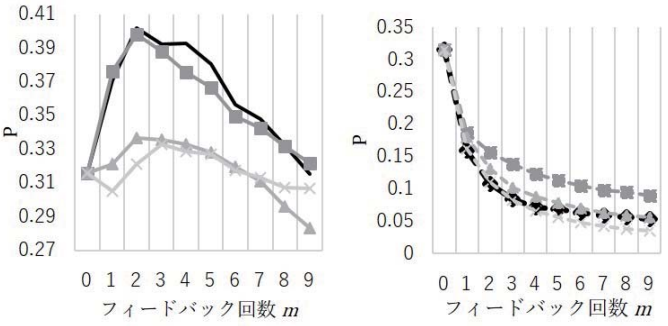

図 5: TF-IDF における検索性能 P

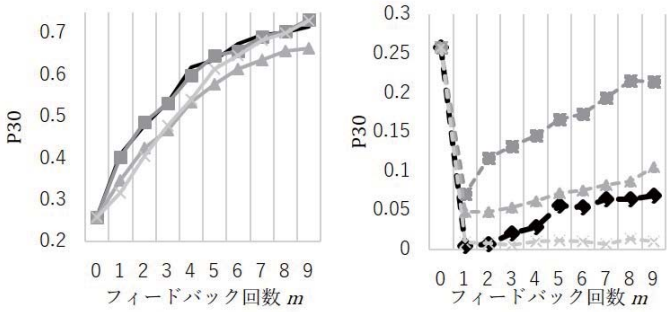

図 6: TF-IDF における学習性能 P30

り、SVM(reduction non-relevant) は提示文書のうち、非適合文書で使用されている属性のみを用いる手順のグラフである。SVM(random) は属性をランダムに選択し、探索空間を縮小した SVM に基づく適合フィードバックのグラフである。ここで、random10000 は、属性をランダムに 10000 個選択したことを示している。SVM(rendom increase) は各フィードバック回において、属性をランダムに選択し、探索空間を広げていく手順を示している。SVM(reduction) と同様の探索空間の大きさで検索を行うように各フィードバック回で増やす属性数を決めている。これらの図より、提示文書全てで使用されていた属性のみを用いる手順により探索空間を縮小した SVM に基づく適合フィードバックでの評価指標の値が、最終的なフィードバック回では、どの条件においても最も高い値を示していることがわかる。ランダムに属性を選択した場合の性能が、探索空間を縮小していない通常の SVM に基づく適合フィードバックと比べ大きく下がってしまうことがわかった。特に、選択する属性数が 1000 の場合では、ほとんど適合文書を発見できていない。通常の SVM と比較し、提示文書全てで使用されている属性のみを用いる手順により探索空間を縮小した場合、最終的なフィードバック回での P の値は TF では約 0.07、TF-IDF では約 0.01 向上した。これは提示した 100 文書のうち、それぞれ平均して TF では 7 文書、TF-IDF では 1 文書、通常の SVM よりも多くの適合文書を発見できていることを示している。

5. 考察

ランダムに属性を選択する手順では、適合フィードバックの性能が大きく下がってしまった。今回の実験のような文書検索では、全ての属性の数に対し、適合文書を判別するための有用な属性の数は非常に少ないと考えられる。このため、ランダムに属性を選択する手順では、適合・非適合の判別に有用な属性を能動的に選択することはできず、選ばれた属性のほとんどが、検索に無意味な属性になってしまうため、性能が大きく下がったと推測することができる。

一方、提示文書のうち、適合・非適合文書で使用されていた属性のみを用いる手順により探索空間を縮小した場合、評価指標は最終的なフィードバック回でどの条件においても最も高い値を示している。この手順は、「システムが適合と判断をしたが、実際は非適合であった文書 (非適合文書)」と、「システムが適合と判断をし、実際に適合であった文書 (適合文書)」に使われていた属性のみを用いることで、適合・非適合の判断に有用であろうと考えられる属性を選択していくものである。ランダムに属性を選択した場合と比べ、能動的に検索システムが分類に必要な属性を選択していくため、多数の検索に無意味な悪い影響を与える属性を除外することができる。このため性能が向上したと考えることができる。

提示文書で使用されている属性のみを用いる場合と、非適合文書で使用されている属性のみを用いる場合では、先ほどの適合・非適合文書両方で使われている属性を用いる手法と比べ、評価指標の値は低くなった。このことより、適合・非適合文書それぞれに用いられている属性の中に、適合フィードバックの性能向上に有効な属性が含まれていることがわかる。

2 種類の探索空間縮小手順の比較により、ランダムに選択した空間では、性能が著しく落ちてしまうが、検索に有用な属性を選択していくことで適合フィードバックの性能を向上させることが可能であることを示すことができた。

TF と TF-IDF による提示文書で使用されていた属性のみを用いる手順により探索空間を縮小した場合の結果を比較すると、TF において性能の向上が TF-IDF より大きかった。これは、TF では単語の頻度のみを重みとしているので、TF-IDF のように、全文書中でのその語の重要度を考慮していない。そのため、通常の SVM に基づく適合フィードバックでは、TF-IDF と比べ検索に無意味な属性の影響をより強く受けてしまと考えることができる。提示文書全てで使用されていた属性のみを用いる手法で探索空間を縮小することにより、この無意味な属性を除くことができていると考えられる。このため、TF の文書ベクトル表現において探索空間を縮小することは、TF-IDF と比べ、適合・非適合文書に含まれていない属性を除くことによる効果が大きかったため、より性能を向上させることができたと考察することができる。

6. まとめ

本研究では、SVM に基づく適合フィードバックにおける探索空間の縮小が性能に与える影響を実験的に分析した。ランダムに属性を選択する手法と、提示文書で使用されていた属性のみを用いる手法の 2 種類の探索空間縮小手順を比較し、TF と TF-IDF の 2 つの文書ベクトルの違いを考慮した実験を行った。その結果、ランダムに属性を選択する手法では性能が著しく落ちてしまったが、提示文書全てで使用されていた属性のみを用いる手法では、特に TF の文書ベクトル表現において、最終的なフィードバック回での性能が大きく向上した。TF において、探索空間を選択することは適合フィードバックの性能向上にとても有効であることが確認できた。検索に有用な属性を選択できる手法であれば、探索空間を縮小することで SVM に基づく適合フィードバックの性能を向上させることが可能であると示すことができた。

参考文献

[1] 村田博士, 小野田崇, 山田誠二:SVM を用いた対話的文書検索における適合性評価の比較分析, 知能と情報 (日本知能情報ファジィ学会誌)Vol23,No.6,pp.853-862(2011)

[2] 小野田崇:サポートベクターマシン, オーム社,(2007)

[3] 黒岩真吾, 柘植覚, 獅々堀正幹, 任福継, 北研二:Simple PCA を用いたベクトル空間情報検索モデルの次元削減, 電気学会論文誌 C 巻：125 号 11pp.1773-1779(2005)