1E1-02

文書検索における非適合性フィードバック手法の一検討

A Study on Method of Non-relevant Feedback for Document Retrieval

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)

The non-relevance feedback method retrieves relevant document using only information on non-relevant documents. We propose a non-relevance feedback method selects presentation documents after deciding the candidate of relevant documents. In this report, we show method of candidate decision and experimental results.

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．この適合性フィードバックにおいて，ユーザがすべての文書を適合しない（非適合）と判定する場合が考えられる．この場合，適合性フィードバックの基本的なアルゴリズムでは，適合文書をもとにして，さらに適合性の高い文書を検索するため，フィードバックが有効に機能しない．この問題を解決するため，我々はユーザが与える非適合文書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非適合性フィードバックを提案している [2].

しかし，提案した非適合性フィードバック手法を大規模な文書データに適用したところ，初期検索で得られた順番に提示していく場合と比較して，明確に良くなる場合と，より悪くなる場合に分かれる結果となった．そこで，本稿では，非適合性フィードバックの新たなアルゴリズムを提案し，その基礎実験結果について報告する．

2. 既提案手法とその問題点

2.1 非適合性フィードバックの概要

非適合性フィードバックは，クエリに対する初期検索結果について，ユーザが非適合の判定しかできなかった場合に，その非適合情報を利用して適合文書を検索する方法である．

その実行ステップは次のようになる．

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，通常の適合性フィードバック検索へ移行する．**

Step 3:非適合性フィードバックの実行

ユーザが判定した非適合文書を用いて，非適合性フィードバックのアルゴリズムから選択された N 文書をユーザに提示して，**Step 2** へ戻る．

2.2 以前提案した非適合性フィードバック手法

我々は，非適合文書を用いて提示文書を決定するため，次のような仮定をおいた．

・適合文書は非適合文書から形成される領域の外部に存在する．
・適合文書は上記仮定を満たす中で非適合文書と似通っている．

一つ目の仮定は，非適合文書に共通する内容の文書は適合文書ではないことを意味し，二つ目の仮定は，非適合文書が初期検索で上位にきていることから，大きくはずれていないことを意味している．この仮定をもとに，与えられた 1 クラスの領域境界を明確化できる One-class SVM[3] を用いた次のような非適合性フィードバック手法を提案した．

1. 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．
2. 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する．
3. **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する**．

2.3 検索実験結果とそこから明らかになった問題点

この非適合性フィードバック文書検索手法の有効性を検討するため，文書検索に関する国際会議 TREC [*1]の第 6 回から第 8 回の adhoc タスクで使用された約 53 万の新聞記事からなるデータセットを用いた実験を行った．その結果，適合文書を早く提示できる割合は提案手法が高かったが，ベクトル空間モデルに基づく文書ベクトルと初期クエリベクトルとのコサイン距離で決定される順位で未判定文書を提示した「フィードバックなし」が適合文書を早く提示できる場合も多数存在した [4].

提案手法は，非適合文書に似通った文書を提示するため，非適合文書と適合文書は「当たらずとも遠からじ」という関係が

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

∗1 http://trec.nist.gov/.

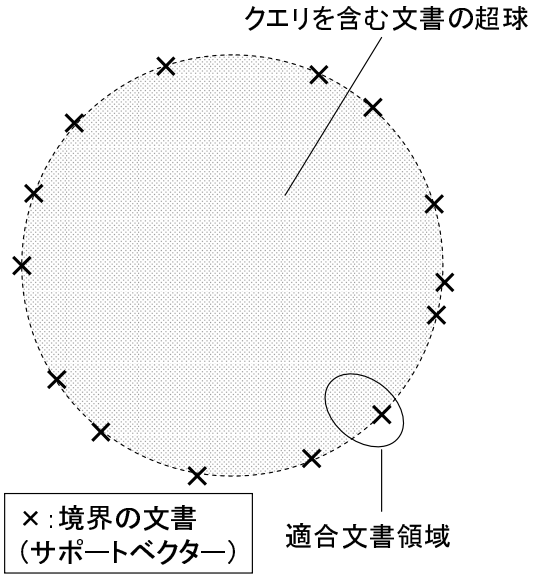

図 1: 提案方式の模式図

あることを前提としている．しかし，非適合文書が適合文書と異なる場合も想定される．たとえば，非適合文書同士が非常に似通っている場合，提案手法は外れた文書を提示し続けることになる．特に，フィードバック文書数 N が小さい場合は，非適合文書のバリエーションが少なくなりやすいため，その傾向が顕著に現れる．

3. 提案する非適合性フィードバック手法

前章で述べた問題点を解決するには，できるだけ内容の異なる文書を提示したほうがよい．適合文書らしい，内容の異なる文書を見つけ出すために，次のような仮定をおく．

・適合文書候補はクエリを含む．
・適合文書候補は他の文書と異なる単語を含む．

一つ目の仮定は，適合文書が少なくともクエリの単語を含んでいるとして， 対象範囲を限定するものである．二つ目の仮定は，前章で述べた問題点への対応として，できるだけ内容の異なる文書を提示するためのものである．文書の内容は，クエリの他にどのような単語を含んでいるかで変化する．他の文書と違う単語を含んでいれば，内容の異なる文書となる可能性は高くなる．

One-class SVM は，大量のデータの中から他と異なる特徴をもつデータで境界となるサポートベクター (以下 SV) を形成する．この性質を利用して，大量のデータからの外れ値検出に利用されている．これは，二つ目の仮定に一致するため，既提案方法とは異なる方法で One-class SVM を適用した非適合性フィードバック手法を提案する．

方式の模式図を図 1 に示す．一つ目の仮定から，クエリを含む文書群が設定され，それに対して One-class SVM を適用すると，文書群の超球の境界に，二つ目の仮定と一致する文書が抽出される．これらの文書が，適合文書になる場合，図のように適合文書領域に含まれる形となる．

4. 提案手法の適用実験

前述の TERC のデータセットを用いて実験を行った．このデータセットは約 53 万の新聞記事文書からなる．TREC の adhoc タスクでは，各回で 50 個ずつの検索課題（トピック）と各課題に適合する文書の情報が提供されている．各トピックは，その内容を 2，3 語で表した title タグ，詳しく記述した description タグ，さらに詳しい適合条件などを記した narrative タグからなっている．本研究では，title タグの単語の組合せをクエリとして使用し，非適合となるクエリができるだけ多くなるようにした．

表 1: 実験結果

フィードバック文書数	適合文書含有率	平均境界文書数	クエリ含有平均文書数
5 文書	91.8%	1940	14218
10 文書	90.6%	2158	16153
15 文書	89.8%	2294	17625
20 文書	89.4%	2383	18571

文書ベクトル表現は TFIDF[2] を用いた．One-class SVM の学習には LibSVM[*2]を使用し，カーネルには RBF カーネルを使用した．RBF カーネルのパラメータ γ は，すべての文書が SV になるのを避け，かつ，ハードマージンになるように調整を行い，1e-3 とした．

実験結果を表 1 に示す．ここで，対象とするクエリは，初期検索のフィードバック文書数 N 中に適合文書が含まれないものだけとしている．したがって，フィードバック文書数の変化により，対象となるクエリが変化する．また，表中の適合文書含有率は，全クエリに対して境界文書中に適合文書が存在したクエリの割合である．表から，平均的に見ると，クエリを含む約 15000 文書を，境界となる約 2000 文書に絞り込んでも，約 90%のクエリについては，境界文書中に適合文書領域が重なっていることがわかる．つまり，提案手法により，適合文書を含むように探索空間を狭めることができる．

5. おわりに

本稿では，新たな非適合性フィードバック手法として，One-class SVM が，大量のデータの中から他と異なる特徴をもつデータで境界を形成する性質を利用して，境界文書を適合文書の候補として提示する方法を提案し，実験を行った．その結果，この方法により適合文書を含むように探索空間を狭められる場合が全クエリの約 90%と高い割合で存在することがわかった．

参考文献

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

[2] 村田 博士, 小野田 崇, 山田誠二: 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2004 年度人工知能学会全国大会論文集:2F1-01, (2004).

[3] 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).

[4] Murata, H. Onoda, T. Yamada, S.: Non-relevance Feedback Document Retrieval using Large Data Set, ISIS 2007:TA02-4, (2007).

∗2 http://www.csie.ntu.edu.tw/cjlin/libsvm/.