

スニペットに対する不適合判断による全文閲覧省略を許す適合性フィードバック

金子 弘明 梅澤 猛 大澤 範高

千葉大学大学院 融合科学研究科 情報科学専攻

1. はじめに

WWW 上で情報検索を行う際、専門性・具体性に乏しい入力クエリに対して目的情報に関連性の低い情報が大量に出力されるために、検索効率が低下するという問題が発生している。しかし洗練されたクエリを作成するには目的情報への深い知識と高い検索技術を要し、困難さを伴う。

この問題解決のために、ユーザの検索クエリ作成支援をする技術として適合性フィードバックがあげられる。これは初期クエリに対して得られた検索結果の目的情報への適合度をユーザに評価してもらい、そのフィードバックを基に検索クエリを修正するという技術である。しかし適合性フィードバックの性能はフィードバックされる文献数が多いほど高くなる傾向にあり[1]、良い結果を得るにはユーザに多くの負担をかける。

このユーザの負担を軽減させる手法として、全文を閲覧せずスニペットのみに基づいて文献の適合度を評価するというものがあるが、この場合本文を閲覧した場合と比較してユーザが評価を誤る可能性がある。そこでこの誤評価の影響を適合度付文書セット NTCIR web-4 を用いて検証するとともに、誤評価を考慮したスニペット利用手法を提案する。

2. 関連研究

陳によればスニペットのみ閲覧した場合と本文全体を閲覧した場合を比較した場合、閲覧時間は60%強削減され、スニペットと本文の適合度評価には10%程度の違いしか見られない[2]。従って、スニペットの利用はユーザの負担を軽減するのに有効であると考えられる。本研究の目的は、スニペットと全文それぞれに基づく適合度評価の差異が実際に適合性フィードバックの性能にどれだけ影響を及ぼすのかを明らかにし、ユーザ負担軽減におけるスニペット利用の有効性を検証することである。

適合性フィードバックの一般的な手法に、適合文献群・不適合文献群の単語情報から検索クエリを拡張する Rocchio アルゴリズムがある。

まず単語出現頻度を要素とする特徴量空間を想定する。各 web ページ・検索クエリは出現単語から

特徴量空間上に点として表現される。検索結果での順位はこの空間上での検索クエリとの距離の題昇順に相当し、検索は距離計算とそのソートに相当する。

Rocchio アルゴリズムは検索結果上位 N 件に含まれる適合文献群 T_r と不適合文献群 T_{nr} を用いて、検索クエリ q_m から次検索クエリ q_{m+1} を生成する。

$$q_{m+1} = \alpha \cdot q_m + \beta \frac{1}{N_r} \sum_{t_{ri} \in T_r} t_{ri} - \gamma \frac{1}{N_{nr}} \sum_{t_{nri} \in T_{nr}} t_{nri}$$

ここで N_r, N_{nr} はそれぞれ検索結果 N 件に含まれる適合文献数と不適合文献数を表し、第2項・第3項はそれぞれ適合・不適合文献群の重心ベクトルを表す。また α, β, γ は任意の重み係数である。

このアルゴリズムにおいて、スニペットのみ閲覧し全文を閲覧しなかった場合に発生しうる誤評価は適合・不適合文献群に含まれる文献の組み合わせに影響し、文献群それぞれの重心ベクトルと次検索クエリを変化させる。

3. 誤評価の影響検証

誤評価の影響検証をするために NTCIR web-4 中の検索課題からそれぞれランダムに 500 件の文献を抽出したものを検証用データセットとした。

それぞれ 500 件中 20 件を初期検索結果とし、Rocchio アルゴリズムにより生成したクエリによる再検索結果と比較して検索性能の向上・低下を測定した。この時、ユーザの評価がある確率で誤るものとして適合・不適合文献群にそれぞれ含まれる文献の組み合わせを変え、結果の変化を比較した。

結果を図1に示す。図中の横軸は誤評価の発生確率を表す。縦軸は初期クエリに対する結果と比べて修正クエリに対する結果がどれだけ向上したかの比を表している。検索性能は次に示す平均精度 (AP: Average Precision) で評価した。また R は適合文献を不適合文献とのみ、 N は不適合文献を適合文献とのみ、 RN は区別なく誤評価が発生した場合をそれぞれ表している。

$$AP = \sum_{i=1}^{20} R(t_i) / i$$

ここで $R(t)$ はユーザが文献 t に対して与える適合度であり、 $\{1, 2, 3, 4\}$ のいずれかの値をとる。適合度が高いほど目的情報により適合しているといえ、それぞれ NTCIR web-4 における C:不適合、B:部分適

Omission Reading Full-text of Non-relevant Document with Snippet Relevance Feedback
Hiroaki KANEKO, Chiba university graduate school of advanced integration science, Osawa Umezawa Lab.

合、A:適合、S:高適合という評価に相当する。また i は検索結果での順位を表す。

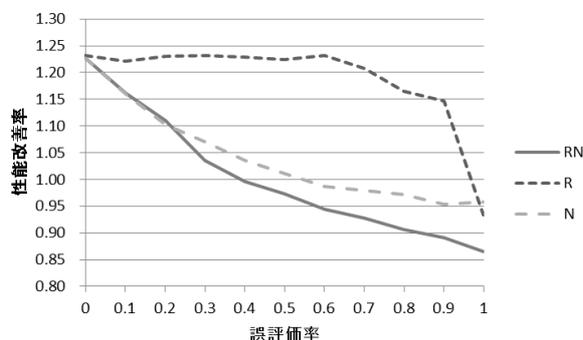


図 1 誤評価による影響

図 1 の R に注目すると、多少の適合文献を誤って不適合文献群に含めてしまっても適合性フィードバックの性能には大きく影響しない事がわかる。また N より、逆に不適合文献を適合文献群に含めると性能が低下する事がわかる。

よって、適合文献群には不適合文献がまぎれていないか精査する必要がある、不適合文献群にはその必要がないと言える。

4. 不適合文献の本文閲覧省略

3 の結果より、適合文献群に分類する際にはよく精査し、不適合文献群に分類する際には簡略的にみることにより、性能を維持したまま閲覧時間を削減することができる。よって、ユーザにはまず文献のタイトルとスニペットのみを提示し、適合と思われるものはさらに本文を閲覧して、不適合と思われるものは本文の閲覧を省略してフィードバックを得るという手法を提案する。

5. 実験

本提案の検証のため、①全文の全文閲覧をした場合、②不適合文献のみ全文閲覧を省略した場合、③適合文献のみ全文閲覧を省略した場合、④全文の全文閲覧を省略した場合での適合性フィードバックの性能とそれに要した時間を比較する。

使用するデータセットは 3 の検証と同じく、NTCIR web-4 の検索課題に基づき、ランダムに 500 件抽出して作成した文献集合を用いる。

初期クエリに対して得られた上位 20 件をユーザに提示し、まずタイトルとスニペットのみから適合度を評価してもらう。次に各文献の全文を提示し、改めて適合度を評価してもらう。この時、提示から評価をフィードバックするまでの時間もあわせて計測する。そして得られたフィードバックから①～④それぞれの場合に相当する適合・不適合文献群を生成し、Rocchio アルゴリズムを用いて修正したクエリで再度検索をかける。この時の検索結果上位 20

件を性能評価の対象とする。評価の指標には AP を用いる。

なお、本実験ではユーザの経時的な知識・興味の変化やそれによる適合評価基準の変化は考慮せず、提示の順番によって評価が変化しないものとする。

6. 結果

結果を表 1 に示す。ここで閲覧時間は 20 件の評価全体にかかった時間であり、性能比は初期検索結果に対する性能向上を AP ではなかったものである。

表 1 全文閲覧省略による性能

	閲覧時間 [s]	性能比
両方ともに全文閲覧	552.3	1.75
適合文献のみ全文閲覧	310.7	1.71
不適合文献のみ全文閲覧	325.7	1.60
両方ともにスニペットのみ	84.0	1.51

なお誤評価は適合・不適合文献群ともに高々 1 割程度発生した。それによって不適合文献のみ全文閲覧した場合・全く全文閲覧を行わなかった場合の性能比が他の 2 つの場合と比較して 1 割程度落ちたという結果は図 1 に見られる傾向と一致する。ちなみに図 1 に対して、②は R、③は N、④は RN にそれぞれ対応する。①は誤評価率 0 における 3 線の交点に対応する。

これより、スニペットから適合文献と思われるもののみ全文を精査し、不適合文献と思われるものは全文閲覧を省略するという提案手法は、適合性フィードバックの効果を低下させずに必要な閲覧時間を削減できる事がわかる。

7. おわりに

適合性フィードバックにおいてユーザ負担軽減の為、スニペットを利用し文献閲覧時間を削減することを考え、それによって発生しうるフィードバックの誤りの影響を NTCIR web-4 を利用しシミュレートし、適合文献を不適合文献と誤っても性能に影響しないことの示唆が得られた。また、適合文献候補のみ精査し不適合文献候補は全文閲覧を省略する試行実験では、性能を維持したまま閲覧時間全体を削減されることを示した。

8. 参考文献

- [1] C. Buckley, G. Salton, J. Allan “The Effect of Adding Relevance Information in a Relevance Feedback Environment” Proceeding of the 17th annual international ACM SIGIR conference on R&D in IR pp.292-300 (1994)
- [2] 陳 明煌, 山田 誠二, 高間 康史 “最小ユーザフィードバックのためのスニペット生成に関する考察” 第 23 回人工知能学会全国大会 2B2-1 (2009)