D-045

# 不適合情報を利用した情報検索手法
# Informatin Retrieval Method using Non Relevancy Information

松村 敦 † 宇陀 則彦 †
MATSUMURA Atsushi UDA Norihiko

## 1. はじめに

利用者が検索システムに投入する問合せの裏側には，その表現からは推測しきれない複雑な検索要求がある．これらの情報を利用できないことが，検索の精度を落す要因の 1 つであることは明らかである．しかし，一般的な利用者は検索システムに対して数語程度のキーワード入力しか行なわない [3]．利用者がこのような行動をとる背景には，複雑な検索要求を容易に入力するインタフェースが存在しない，複雑な検索要求を十分に検索結果に反映させる手法がない，といった問題がある．いずれにしても，情報検索システムには数語程度のキーワード入力から精度の高い検索結果を出せるだけではなく，より複雑な検索要求を理解しそれを検索結果に反映させることができるような仕組みが求められる．

従来から問合せの情報不足を補う手法としては，シソーラスや疑似適合フィードバックを利用した質問拡張が提案され盛んに研究されている [5]．しかしながらこれらの手法は，あらかじめ与えられた共通の知識あるいは検索された文書集合によって自動的に情報を補っているため，個々の利用者にとって本当に最適な手法であるとは言い難い．

これに対して，筆者は利用者の入力した自然言語文による問合せを係受け解析し，キーワードの集合から欠落してしまう機能語の役割を利用した検索手法を提案してきた [7]．この手法によって利用者自身の意図を検索に反映することが可能となったが，一方で，さらなる検索精度の向上のためにはいくつかの特徴的な情報を個別に正確に利用する必要があることが明らかとなった．例えば，否定語の否定する内容を正確に把握することが検索精度向上に大きく関わってくる．

このような背景から，本研究では利用者の持つ複雑な検索要求の中から検索精度に大きく影響する要素を明らかにし，これを利用することによって利用者の検索意図をよりよく反映する検索手法の開発を行なうことを目的とする．ここで，本研究では利用者が複雑な検索要求を持つ (あるいは持つようになる) 場合を想定するが，今回は具体的な利用者の要求として，情報検索システム評価用テストコレクション NTCIR-1 (本格版) および NTCIR-2 (本格版) の検索課題を対象とした．

以下，2. 節では NTCIR テストコレクションの検索課題の分析を行なう．これをもとに 3. 節ではその中から抽出した「不適合条件」の構造化とこれを利用した検索手法について示す．4. 節では実際の検索実験とその結果について述べ，5. 節で結論と今後の課題について述べる．

†筑波大学大学院図書館情報メディア研究科

## 2. 検索要求の構造

情報検索システム評価用テストコレクション NTCIR-1 (本格版) および NTCIR-2 (本格版) [1] の検索課題には TITLE, DESCRIPTION, NARRATIVE というフィールドがある．これらは検索要求をそれぞれ 1 語程度，1 文程度，複数文の詳しい文章，という形式で検索要求を持つ本人によって記述されたものである．本研究では，この NARRATIVE の情報を利用者の持つ複雑な検索要求と位置付け分析対象とする．

これまでの NTCIR を利用した研究を見ると，TITLE や DESCRIPTION だけでなく NARRATIVE も利用した方が検索精度が高くなるという結果が出ている．これは，主に関連するキーワードが増えることにより検索要求がより明確になるということによるものである．しかしながら，NARRATIVE に書かれているような複雑で詳しい説明文を検索システムに入力するよう求めることは利用者の負担という点からまず考えられない．したがって，NARRATIVE の中の重要な要素のみを利用して利用者の負担を押えつつ，検索精度の向上を狙う必要がある．このような観点から NTCIR-1 の検索課題 83 件と NTCIR-2 の検索課題 49 件を対象に分析を行なった．

その結果，NARRATIVE は「背景」「適合条件」「不適合条件」の 3 つの内容から成ることが分かった[2]．「背景」は検索要求が生じた理由や，その分野の歴史などである．「適合条件」は，適合であるための追加の情報や詳しい条件である．「不適合条件」は適合でないものに対する条件を提示しているものである．

一般に，「背景」や「適合条件」は検索要求の曖昧性を解消し検索結果を絞り込む効果と，逆に，関連する情報によって検索結果を広げる効果の両方を持つ．これに対して，「不適合条件」の場合は主に前者の効果を持つ．これは否定表現を含むことによる大きな特徴である．そこで，本研究では利用者の検索要求の 1 つとして不適合条件に着目し，これを構造化し検索に利用する方策を検討した．文末表現を手がかりとして検索課題の NARRATIVE から抽出された不適合条件文は 110 文 (NTCIR-1 に 68 文，NTCIR-2 に 42 文) である．

## 3. 不適合条件を利用した検索手法

利用者の入力した不適合条件を検索に利用するためには，不適合を表す内容 (不適合内容) を正確に表現し，こ

[1] http://research.nii.ac.jp/ntcir/

[2] NTCIR-4 からは NARRA (NARRATIVE と同等のタグ) の文に BACK(背景)，REL(適合情報)，TERM(キーワード) というタグを付与し，あらかじめ役割が示されるようになった．ただし，REL には適合，不適合の両方の情報が区別なく書かれている．

| 不適合型 | 不適合条件文の例 |
|---|---|
| 肯定 | インスリン注入型人工膵移植 [は不可。] |
| 否定 | 特異点の必然性について言及していない**もの** [は不可。] |
| 以外+肯定 | B 型肝炎**以外の**ワクチン [も不可。] |
| のみ | 宇宙定数の測定方法や測定計画について**のみ**述べたもの [は、要求を満たさない。] |
| 以外+肯定+以外+肯定+のみ | 小細胞癌**以外**の組織型や肺**以外の**部位に関して**のみ**論じているもの [は不可。] |

表 1: 不適合表現の分類

れを検索に反映させる必要がある．以下では本研究で提案する不適合条件を利用した検索手法について説明する．

はじめに，不適合条件を「は不可とする。」や「は適合しない。」のような文末表現と不適合の条件を表す不適合内容 (名詞句) とに分割する．分割には正規表現による文末規則を作成し，これを利用した．不適合内容は不適合条件を表す本質的な部分である．

不適合内容は以下の 4 種類の要素から構成される．これを不適合要素と呼ぶ．

**肯定** 肯定表現になっている部分

**否定** 否定表現になっている部分

**以外** 「以外」が含まれる部分

**のみ** 「のみ」が含まれる部分

各不適合内容は上記 4 種類の不適合要素のいずれか，またはその組合せによって表現される．この組合せを不適合型と定義する．各型に分類された不適合条件文の例を表 1 に示した．表中の例では，不適合文末表現は括弧で括り，分類の指標として利用した特徴語をゴシック体で示している．これらの型は検索要求と不適合内容との意味関係を反映したものである．後に述べるように検索の際にはこれらの型に応じ検索手法を変えることになる．

このような不適合型を同定するために不適合条件の係受け解析を行なう[3]．係受け解析の結果から，不適合文末表現を削除し，「ない」，「以外」，「のみ」等の特徴語に係る単語の集合を同定する．以上の処理により，最終的に不適合要素を [特徴] { 単語集合 } で表現し，これらの組合せで不適合型を同定し，不適合条件を表現する．例えば，「B 型肝炎以外のワクチンも不可。」は以下のように表現される．

[以外] {B, 型, 肝炎} [肯定] {ワクチン}

[3]係受け解析には CaboCha を利用した．
http://chasen.org/~taku/software/cabocha/

| 不適合型など | $\alpha$ の設定 |
|---|---|
| 肯定/のみ | -0.5 |
| 否定/以外 | +0.5 |
| のみ + 肯定 | それぞれ-0.5 と +0.5 |
| 否定/以外 + 肯定/のみ | それぞれ+0.5 と +0.5 |
| 肯定+否定 | それぞれ-0.5 と +0.5 |
| 並列の場合 (や/、/など) | 肯定/のみには-0.5，否定/以外には+0.5 |

表 2: 各不適合型に対する $\alpha$ の設定

以上のようにして表現した不適合条件を検索に利用する手順は以下の通りである．

1. 初期検索を行ない各文書 $d$ に対する初期文書得点 $SI_d$ を求める．
2. 各不適合要素 $i$ に含まれる単語集合による文書得点 (不適合要素得点) $SN_{d,i}$ を検索結果の各文書に対して求める．
3. 初期文書得点と不適合得点を不適合表現の型に応じて組み合わせて総得点 $S_d$ を求める．
4. 総得点で文書をランキングし出力する．

手順 3 の具体的な得点計算式は式 (1) で示す単純な線形結合の式で定義した．

$$S_d = SI_d + \sum_{i=1}^{N} \alpha_i \times SN_{d,i} \quad (1)$$

ここで，$SN_{d,i}$ の計算の際には，要素に含まれる単語のうち，「研究」「論文」「もの」といった一般的な語は省く．また，初期検索に含まれる単語も同様に省いて計算する．初期検索に含まれる単語は $SI_d$ で既に考慮されているため，再度，得点要素として加算することはその語の持つ重みを過剰に利用することになり，適切な文書得点を与えることができなくなるからである．

$\alpha_i$ は不適合要素 $i$ の性質に応じて得点を調整するためのパラメタである．絶対値を 0.5 に固定して予備実験を行ない符号を決定した．これらを表 2 にまとめた．

不適合内容が肯定されるか否定されるかを考慮し，基本的に肯定的な要素に対しては負の重みを，逆に否定的な要素に対しては正の重みを与える．要素の組合せや列挙があった場合にはこれらを加算することで複数の要素の効果を求める．

## 4. 検索実験と評価

本手法による不適合条件の利用の効果を検証するために NTCIR-2 (本格版) を利用して実験を行なった．NTCIR-2 の検索対象は約 73 万件の学術文書の抄録，検索課題は

| 手法 | MAP | 向上率 |
|---|---|---|
| baseline (BM25) | 0.1661 | – |
| Rocchio 型フィードバック | 0.2035 | 22.5% |
| 本手法 | 0.2031 | 22.3% |

表 3: 平均適合率の平均 (TITLE のみの場合)

| 手法 | MAP | 向上率 |
|---|---|---|
| baseline (BM25) | 0.3156 | – |
| Rocchio 型フィードバック | 0.3321 | 5.2% |
| 本手法 | 0.3118 | -1.2% |

表 4: 平均適合率の平均 (DESCRIPTION のみの場合)

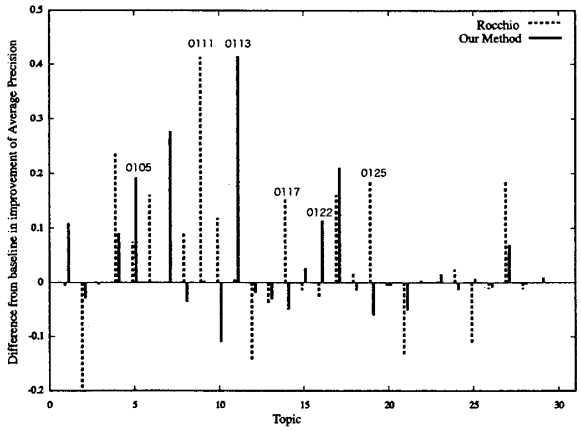


図 1: 検索課題毎の平均精度の baseline との差 (TITLE のみの場合)

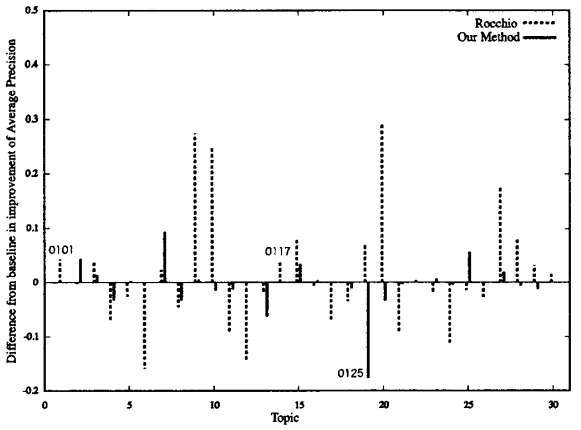


図 2: 検索課題毎の平均精度の baseline との差 (DESCRIPTION の場合)

49 件である．この中から今回は不適合表現を含む 30 件の検索課題を利用して評価を行なった．適合判定は，高適合 (S)，適合 (A)，部分的適合 (B)，不適合 (C) の 4 段階判定で行なわれているが今回の実験では S と A を正解とした．初期検索は短い検索質問を想定するため，検索課題の TITLE のみを用いた場合と DESCRIPTION のみを用いた場合について実験を行なった．

基本となる検索システムには情報検索パッケージ [6] を利用した．このパッケージは BM25[4] を採用し，NTCIR-2 テストコレクションでは上位のシステムと同等の検索精度を達成している．また，同パッケージに含まれる Rocchio 型の疑似適合フィードバック [1] (以下，Rocchio 型) を利用した検索も行ない比較の対象とした．

はじめに，TITLE のみを利用した場合の 3 つの手法の平均精度の平均 (Mean Average Precision, MAP) を表 3 に示す．この結果から分かるように，baseline に比べて Rocchio 型は 22.5%，本手法は 22.3%の精度向上を達成している．わずかに Rocchio 型の方が精度が高いがほぼ同等であるといえる．

しかしながら，検索課題毎の平均精度は Rocchio 型と本手法では大きく異なっている．図 1 に問合せ毎の平均精度の baseline との差をプロットしたものを示す．この図から，非常に多くの問合せで両手法の精度が大きく異なっていることが分かる．具体的には半数以上の 16 件の問合せで両手法の精度は 0.1 以上の開きがある．これは，Rocchio 型と本手法の効果は別々の要因でおきており，両手法が相補的な関係にあることを示している．

検索課題を個別に見ると，例えば検索課題 0122 では，Rocchio 型が baseline に比べ精度を落しているのに対して，本手法は大きく精度を上げている．この検索課題の TITLE は「リテラシーと教育」であるが，不適合条件は「情報リテラシー，メディアリテラシー，コンピュータリテラシーなどは含まない。」となっている．本手法は，TITLE の「リテラシー」という単語で検索されてしまうこれらの概念を含む文書を的確に排除することに成功している．

また，否定や以外の不適合要素を含む場合は結果として検索質問の適切な拡張を行なっていることも確認できた．検索課題 0105 では，TITLE が「新規キノロン剤」であるのに対して，不適合条件は「ウサギ、ラット以外の動物を用いたものは不可。」である．この場合は本手法によってウサギ，ラットを含む文書の得点を上げることに成功し，精度も向上している．

一方，不適合条件が適切な質問拡張の邪魔をしている場合には精度が落ちている．例えば，検索課題 0125 の場合には，「電解水」が TITLE，不適合条件が「機能性水と表記されている中で、次亜塩素水や消毒剤の水溶液など電解による生成水でないものは除く。」である．不適合条件に含まれる「次亜塩素水」「消毒剤」は質問拡張に利用することで検索精度が向上するが，本手法ではこれを阻害する方向で扱ってしまう．

次に，初期検索に DESCRIPTION を用いた場合についての結果を表 4 に示す．この場合，Rocchio 型は 5.2%の精度向上を達成しているのに対して，本手法は baseline

よりも精度が落ちている．検索課題毎に baseline からの差を示したグラフ (図 2) を見ても分かるように，本手法ではほとんどの検索課題で精度向上を達成できていない．

本手法が機能しない原因の 1 つに，DESCRIPTION と不適合条件の重なりがある．例えば，検索課題 0101 の場合，DESCRIPTION が「遺伝子工学的手法による B 型肝炎ワクチンの開発について論じている文献」であるのに対して，不適合条件は「遺伝子工学的手法に触れていない論文は不可。」と「B 型肝炎以外のワクチンも不可。」である．このように同一の単語で同一の意味を述べている場合には不適合条件は利用されないため baseline からの精度向上はない．また，TITLE の場合と同様に検索課題 0125 においては，質問拡張に効果のある単語を不適合情報として利用することになり，かえって精度を落している．さらに，キーワードの集合で不適合内容を表現する限界を示した例もある．検索課題 0117 の DESCRIPTION は「歴史史料を電子化し、データベースとしてインターネット上で利用できるようにしたものはないか。」であり，一方，不適合条件は「インターネット上で利用できる史料を使って行なった歴史研究などは含めない。」である．両者はキーワード集合で見ればほぼ同等であり，本手法は機能しないが，実際には文の意味は全く異なる．このような場合には，より詳細に不適合条件の内容を獲得する必要があり，不適合要素の表現をキーワード集合ではなく係受け関係とする方法などを検討する必要がある．

## 5. おわりに

利用者の複雑な検索要求の 1 つとして不適合条件に着目し，これを構造化し検索へ反映させる手法を提案した．NTCIR-2 テストコレクションを利用して評価実験を行なった結果，初期検索を TITLE のみで行なった場合の検索精度は Rocchio 型適合フィードバック手法と同等であった．これにより TITLE のみの検索のように初期検索に入力する情報が非常に少ない状態での本手法の有効性が示された．また，検索課題毎に精度を比較すると，本手法と Rocchio 型適合フィードバック手法では全く異なる傾向を示し，互いに相補的な関係にあることが明らかになった．一般に，疑似適合フィードバックはキーワードを補うことで検索漏れに対処することを主眼としているのに対して，本手法は利用者が明示的に示した不適合条件によって不要な結果を排除することを目指している．今回の結果は，このような 2 つの手法の違いを反映するものであり，両手法を適切に使い分けることで，より高度な検索システムを実現できることを示唆している．

一方で，初期検索に DESCRIPTION を用いた場合には，Rocchio 型適合フィードバック手法は性能が向上したが，本手法は精度を上げることができなかった．DESCRIPTION のように 10 語程度の長さの検索質問を与える場合には，BM25 による重み付けと Rocchio 型適合フィードバック手法は適切に機能するが，不適合情報を利用する本手法は効果的に機能しない．

今後は，結果の詳細な分析と得点付け手法の改良をすすめる．現在は不適合要素が複数ある場合には単純に加減で処理しているが，不適合要素の粒度も不均一であるため，より精密な得点付けが必要と思われる．Dkaki らは本研究と同様に複雑な検索要求を構造化し，パッセージ検索への応用を試みており [2]，得点付け手法の改良にはこの手法を参考にする．また，今回は不適合情報に着目して構造化を行なったが，適合情報，背景情報の構造化も行ない，利用者の情報要求をより多く獲得することも課題となる．

今回は，NTCIR の NARRATIVE を利用者の検索要求として扱ったが，現実の検索場面で利用者がこのような形式で要求を持つかどうかは明らかではない．将来的には，インタラクティブな検索も視野に入れて利用者の検索要求についての調査も必要になると思われる．

## 謝辞

本研究では，国立情報学研究所の提供する情報検索システム評価用テストコレクション NTCIR-1 (本格版) および NTCIR-2 (本格版) を利用した．

## 参考文献

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[2] T. Dkaki and J. Mothe. Combining Positive and Negative Query Feedback in Passage Retrieval. In *Proceedings of RIAO2004*, pp. 661–672, 2004.

[3] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: A study of user queries on the web. SIGIR Forum, 32(1), pp.5–17, 1998.

[4] S. E. Robertson and S. Walker. Okapi/Keenbow at TREC 8. In *Proceedings of TREC 8*, pp. 151–162, 2000.

[5] T. Sakai, M. Koyama, A. Kumano, and T. Manabe. Toshiba BRIDJE at NTCIR-4 CLIR: Monolingual/Bilingual IR and Flexible Feedback. Working Notes of the Fourth NTCIR Workshop Meeting, 2004. http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/index.html.

[6] 内山将夫, 井佐原均. 情報検索パッケージの実装. 情報処理学会情報学基礎研究会研究報告, FI-63-8, pp.57–64, 2001.

[7] 松村敦, 高須淳宏, 安達淳. 情報検索における単語間の関係の効果. 情報処理学会研究報告, Vol. 2001, No.70, pp.257–264, 2001.