係受け関係を利用した検索意図の抽出と情報検索への適用

松村 敦
筑波大学 図書館情報メディア研究科
matsumur@slis.tsukuba.ac.jp

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．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ある程度の検索精度の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．

これまで，筆者はキーワード間の係受け関係を考慮することによって，キーワードの集合から欠落してしまう機能語の役割を検索に組み込む手法の研究を行なってきた [4]．その結果，検索の精度を向上させるためには，いくつかの特徴的な係受け情報を正確に利用する必要があることが分かった．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の大きな要因となることなどである．

一方，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 [2]．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えない．

利用者の持つ複雑な検索要求のうち，本当に重要な要素 (これを「検索意図」と呼ぶ) は何であるか．本研究はこのような視点に立ち，利用者自身が複雑な検索要求を持つ (あるいは持つようになる) 場合を想定し，これを活用することによって利用者の意図を反映できる検索手法の実現を目指す．

今回は利用者の発した検索要求の分析から検索意図の 1 つとして「不適合条件」に着目し実際の検索実験を通して効果的に利用可能かの検証を行なった．分析および実験の対象は，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) である[1]．

以下，2 節では検索要求の分析，3 節では検索手法，4 節では実際の検索実験を述べ，5 節で結論と今後の課題を述べる．

2 検索要求の分析

NTCIR-1 および NTCIR-2 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．これまでの NTCIR を利用した研究では，TITLE や DESCRIPTION のみを利用した場合よりも，NARRATIVE を利用した方が一般的には精度が高くなるという結果が出ている．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 ほど詳しくなくとも重要な要素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．

分析の結果，NARRATIVE は主に「背景」「適合条

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

件」「不適合条件」の 3 つの内容に分割できることが分かった[2].「背景」は検索要求が生じた理由，その分野の歴史などの背景情報である.「適合条件」は，適合であるための追加の条件,「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．実際の検索場面では，多くの結果が出てきた場合には意図しない検索結果を省いて再検索したいという要求は多い．また，前節で述べたように否定の関係を利用することが検索精度の向上のために重要であるという結果もある．これらを考慮して，今回は検索意図の 1 つとして不適合条件に絞って検索に利用する方策を検討した．

はじめに，検索課題の NARRATIVE から不適合条件文の抽出を人手で行ない，110 文 (NTCIR-1 に 68 文, NTCIR-2 に 42 文) を得た．これらの不適合条件文を「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) とに分割した[3].

次に，不適合内容の記述形式をもとにして次の 5 種類の型を定義し，全ての不適合条件文を分類した．

肯定型 不適合内容が肯定表現になっている文.

否定型 不適合内容が否定表現になっている文.

以外型 不適合内容に「以外」が含まれる文.

のみ型 不適合内容に「のみ」が含まれる文.

その他 上記に当てはまらない表現および,「〜以外〜のみ〜は不可である。」のように複数の型の組合せとなっている文.

実際の例を表 1 に示した．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太字で示している．この分類は検索要求と不適合内容との関係を念頭に置いて作成したもので，検索の際には型に応じて検索手法を変えることを想定している．

最後に，不適合表現に含まれる単語の役割を同定するため，以下の手順で解析を行なった．

1. 不適合表現を係受け解析する[4].
2. 不適合文末表現を削除する.
3. 単語の係受け情報を利用して「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，最終的に不適合条件を [型名] { 単語集合 } の形で表現する．例えば,「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。」は [のみ]{ 宇宙，定数, 測定, 方法, 計画 } のように表現される.

型	不適合表現の例
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	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。]
その他	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

3 検索手法

前節の手順で表現された不適合内容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. 不適合内容を構成するキーワード集合による文書得点 (不適合得点) SN_d 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.

今回の実験では初期文書得点と不適合得点の組合せ方法に式 (1) で示す単純な線形結合を利用した.「肯定

[2]NTCIR-4 からは NARRA (NARRATIVE と同等のタグ) の文に BACK，REL，TERM 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.

[3]分割自動化のために正規表現による文末規則を作成したが，本研究は不適合条件文の分割自体には焦点をあてていないためここでは言及しない.

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

型」と「のみ型」は不適合内容が否定されるのに対して,「否定型」と「以外型」は不適合内容が肯定されることを考慮し不適合得点の符号を変えている．ここで,$\alpha, \beta, \gamma, \delta$ はパラメタである．

$$S_d = \begin{cases} SI_d - \alpha \times SN_d & (\text{「肯定型」の場合}) \\ SI_d + \beta \times SN_d & (\text{「否定型」の場合}) \\ SI_d + \gamma \times SN_d & (\text{「以外型」の場合}) \\ SI_d - \delta \times SN_d & (\text{「のみ型」の場合}) \end{cases} \quad (1)$$

4 検索実験と評価

検索実験には NTCIR-2 (本格版) を利用した．検索対象は約 73 万件の学術文書，検索課題は 49 件である．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 を正解として評価を行なった．

実験に利用した検索課題は 49 件のうち「肯定型」の不適合表現を含む 12 件,「否定型」の不適合表現を含む 10 件,「以外型」の不適合表現を含む 6 件,「のみ型」の不適合表現を含む 8 件である．初期検索は単純なキーワード検索を想定するため，検索課題の DESCRIPTION を用いた．

基本となる検索システムには情報検索パッケージ [3] を利用した．このパッケージは BM25[1] 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．したがって baseline としては十分なシステムと考えて良い．また，同パッケージに含まれる Rocchio 型の適合フィードバックを利用した検索も行ない比較の対象とした．

最初に，式 (1) のパラメタを 0 から 1 の間で 0.1 刻みで変化させて実験を行ない本手法の精度の変化を調べた．比較対象の baseline および Rocchio 型よりも本手法の精度が高かった検索課題の件数が最も多かった結果をそれぞれ表 2，3，4，5 に示した．パラメタの具体的な値は $\alpha = 0.3, \beta = 0.5, \gamma = 0.5, \delta = 0.4$ である．表の各数値は検索精度を評価する指標の 1 つである平均精度 (Average Precision) であり，各検索課題に対して最も精度の高かったものを太字で示している．

それぞれの結果から，本手法が非常に有効な場合とそうでない場合があることが分かる．さらに詳しく分析するといくつかの特徴的な傾向が見られる．まず，本手法が最も精度の高かった問合せをみる．例えば，検

topic ID	baseline	Rocchio 型	本手法
0102	0.4249	0.4232	**0.4972**
0112	0.3559	**0.6067**	0.2325
0116	**0.4638**	0.4476	0.3140
0117	0.4517	**0.4937**	0.3613
0122	**0.0084**	0.0000	0.0034
0124	**0.6584**	0.6245	0.6317
0127	0.2889	**0.5786**	0.3109
0129	0.1141	0.0227	**0.1809**
0136	**0.2619**	0.1512	0.1907
0141	**0.1384**	0.1117	0.1324
0147	0.0183	**0.0976**	0.0059
0149	0.0074	**0.0239**	0.0073

表 2:「肯定型」不適合表現を含む検索課題の実験結果

topic ID	baseline	Rocchio 型	本手法
0101	0.4002	**0.4439**	0.3274
0104	**0.5577**	0.4877	0.5039
0107	0.3407	0.3639	**0.4607**
0111	0.0292	**0.3034**	0.0463
0113	0.7850	0.6927	**0.8202**
0119	0.1047	**0.1841**	0.1160
0123	**0.5847**	0.5168	0.4512
0125	0.5370	**0.6050**	0.5707
0132	0.0446	0.0254	**0.0511**
0143	0.4952	**0.6696**	0.3141

表 3:「否定型」不適合表現を含む検索課題の実験結果

topic ID	baseline	Rocchio 型	本手法
0101	0.4002	**0.4439**	0.3406
0105	0.2590	0.2344	**0.2957**
0107	0.3407	**0.3639**	0.3259
0114	**0.2241**	0.0822	0.2150
0143	0.4952	**0.6696**	0.5598
0148	0.1479	0.1795	**0.2052**

表 4:「以外型」不適合表現を含む検索課題の実験結果

topic ID	baseline	Rocchio 型	本手法
0102	0.4249	0.4232	**0.5559**
0103	0.1186	**0.1576**	0.1017
0109	**0.4942**	0.4489	0.3839
0112	0.3559	**0.6067**	0.3061
0129	**0.1141**	0.0227	0.0565
0137	0.5073	0.4943	**0.5625**
0147	0.0183	**0.0976**	0.0096
0148	0.1479	**0.1795**	0.1543

表 5: 「のみ型」不適合表現を含む検索課題の実験結果

索課題 0102 では「肯定型」(表 2),「のみ型」(表 5) ともに本手法が有効である．検索課題 0102 の初期検索に利用した DESCRIPTION は「糖尿病治療のための異種膵島移植の長期生着例について論じている文献」である．これに対して，不適合条件は「肯定型」の「インスリン注入型膵移植は不可。」および,「のみ型」の「同種同系移植のみを対象としたものも不可とする。」である．初期検索のキーワードと重複していない重要なキーワードが不適合条件に含まれており，これらを考慮することが検索精度の向上に貢献している．

これとは反対に，本手法が精度を大きく落している例では，DESCRIPTION と不適合条件の言及する内容がほぼ一致している場合が多い．例えば，検索課題 0101 の場合 (表 3，4) ，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「否定型」の「遺伝子工学的手法に触れていない論文は不可。」と「以外型」の「B 型肝炎以外のワクチンも不可。」である．いずれの不適合条件についても DESCRIPTION の内容の言い直しにすぎないため，不要な検索結果の削除に効果はなく，キーワードの重みを過剰に付与し，精度を下げる結果となっている．

また，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に検索課題 0117 (表 2) がある．この場合の 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「肯定型」の「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るが，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を獲得する必要があり，不適合内容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

利用者の検索意図の 1 つとして不適合条件に着目し，これを検索結果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，Rocchio 型のフィードバックと比較して，本手法が有効な場合とそうでない場合があることが分かった．

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洩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文書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，より有効な検索システムを実現できることを示唆している．

今後の課題は，結果の詳細な分析を進めるとともに，不適合条件を反映させる検索手法の改良と不適合条件だけでなく適合条件も利用するための仕組みの検討を行なうことである．

参考文献

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

[2] 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.

[3] 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.

[4] 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.