D-005

全文検索における精度向上

－ 検索語の拡張と削除について －

Precision Improvement in Full-text Retrieval

菅 豊† Yutaka Kan
佐藤 隆士‡ Takashi Sato

1．はじめに

今日では，オンライン検索システムやインターネットサーチエンジンなどに用いられているような，情報検索に関する研究・開発が世界中において進められ，また検索実験におけるプロジェクトが行われている．同様に日本においても検索実験を行うプロジェクトが存在し，その内の 1 つとして現在の国立情報学研究所により開始され，現在も継続中である NTCIR（NII/NACSIS Test Collection for Information Retrieval）[1]プロジェクトがある．

NTCIR の大きな特徴としては，主な検索対象が日本語で書かれたテキストや文書を含むという点である．日本語で書かれたテキストに対し情報検索する場合，未だ解決できていない問題もあり，更なる情報検索技術の発展が望まれている．そのためには，情報検索分野以外の研究者も NTCIR に参加し，多方面から見た情報検索の研究が行われていく必要がある．

NTCIR は現在までに 6 回開催されており，前回行われた NTCIR-6 における成果報告会（ワークショップ）は 2006 年 4 月から 2007 年 5 月までの期間においての成果報告がなされた．

本稿では，前回行われた NTCIR-6 の結果から，検索精度を向上させるために，検索語を拡張・削除における評価により検索語の改良を行った結果を記す．

2．検索実験

2.1 実験概要

検索実験では，いくつかの検索課題を抽出し，それらの課題について検索語を改良・評価することにより，検索精度の向上を図る．今回実施した実験においては，次のようにして検索課題の抽出を行った．

NTCIT-6 の正誤判定結果である，CLIR Stage1 J-J Relax.txt から S，A 判定であった正解の記事を取出し，それらの記事の中から，検索語の検討のしやすさや検索実験の容易さなどを考慮して，正解の記事件数が 15 件以下であったものを選択基準として，検索課題の抽出を行った．その結果，8 題の検索課題が抽出された．

2.2 実験内容

検索実験に使用するために抽出した，いくつかの正解の記事の<TEXT>内から検索課題の DESC，NARR に従い，それぞれの課題に適切なキーワードを検討した上で，検索語を改良し，再び検索実験を行う．ここで，今回の実験における検索語の改良に伴う語の拡張・削減は，手動により行っている．

今回の検索実験における評価では，NTCIR-6 において我々が提出した Run の Average precision の値を比較する．Average precision とは平均精度のことであり，この値の向上は，検索課題に対する検索語の適合度（精度）が向上することを意味している．

2.3 検索課題

今回の実験において抽出した検索課題の 1 つを例として説明する．

```
<TOPIC>
<NUM>070</NUM>
<ONUM>NTCIR3-98-010</ONUM>
<SLANG>CH</SLANG>
<TLANG>JA</TLANG>
<TITLE>逆エルニーニョ現象</TITLE>
<DESC>
逆エルニーニョ現象とは何であるか、および、逆エルニーニョ現象との比較について探す。
</DESC>
<NARR>
地球規模の気象において、エルニーニョ現象の後に続く逆エルニーニョ現象の影響とはどのようなものか？ エルニーニョ現象との比較とはどのようなものか？ 逆エルニーニョ現象についての基本的な紹介、どのようにして起こるのか、その特徴と伝播、などを適合とする。エルニーニョ現象によって引き起こされる、ある特定の国における影響は、不適合とする。
</NARR>
<CONC>
エルニーニョ現象, 逆エルニーニョ現象, 気象
</CONC>
</TOPIC>
```

この例は，逆エルニーニョ現象についての検索課題の例である．ここで DESC，NARR に従い検索語を検討し，その検索語により検索実験を行うことで，このような検索課題に対する検索語の精度を検証する．

3. 実験結果

3.1 検索語の拡張

（1）先の検索課題の例では，拡張語の例として「逆エルニーニョ現象」「エルニーニョ現象」「気象」「特徴」「伝播」などが挙げられる．また抽出された正解の記事から，逆エルニーニョの正式名称が「ラニーニャ」であることも分かる．この例における元々の検索語は CONC（concept）で示されている 3 つの語であったが，上で示したように「特徴」「伝播」「逆エルニーニョ」「ラニーニ

†大阪教育大学大学院 教育学研究科 総合基礎科学専攻 数理情報コース

‡大阪教育大学 情報処理センター

ヤ」といった 4 つの語により検索語の拡張を行い，再び検索実験を行った結果，表 1 のように精度が向上した．

表 1　語拡張の結果

	拡張前	拡張後
Average precision の値	0.0011	0.3542

(2) 抽出された検索課題の例として，もう 1 つ検索課題を見るために，その一部を以下に示す．この例は，ティーンエイジャーのファッションに関する検索課題である．

<TITLE>
ティーンエイジャーのファッション
</TITLE>
<DESC>
服装や髪型、化粧におけるティーンエイジャーのファッションを記述した文書を検索する。
</DESC>
<NARR>
ティーンエイジャーが関心を持つ服装や髪型、化粧などについてのファッション動向についての文書。適合文書は、中学、高校、大学の新入生などのティーンエイジャーについてのものを示していなければならない。服装、髪型、化粧、靴などの姿かたちに関するアイテムのうちの 1 つのみを扱った文書も適合する。もしその文書が商品やその価格の紹介である場合、不適合である。
</NARR>
<CONC>
ティーンエイジャー，X世代，ファッション，服装，髪型，化粧
</CONC>

この例でも元々の検索語は CONC にある 6 つの語であった．ここで拡張語の例として「１０代」「メイク」「ブーツ」「ヒール」「バッグ」など合計 24 の語として再び検索実験を行った結果，表 2 のように精度の向上が見られた．しかし先の例とは違い，大幅な向上は見られなかった．

表 2　語拡張の結果

	拡張前	拡張後
Average precision の値	0.0071	0.0682

(3) 上記の例のように，前途の選択基準により抽出された全 8 題の課題において検索語を拡張し，検索実験を行った結果，表 3 のように，少々精度の向上が見られた．

表 3　語拡張の結果

	拡張前	拡張後
Average precision の値	0.1116	0.1793

3.2 検索語の削減

(1) ここで，1 つ目の検索課題の例において，拡張した検索語の内「気象」「特徴」「伝播」という語は今回の検索課題だけでなく，他の課題にも頻出するような一般的な語であると予想される．つまり，この 3 つの語は検索精度を下げる要因となっていると考えられるため，これらの語を除いた 4 つの語「逆エルニーニョ現象」「エルニーニョ現象」「逆エルニーニョ」「ラニーニャ」で検索実験を行った結果，Average precision の値は 0.5833 となり，検索精度が向上した．

(2) 先ほどと同様に，「ファッション」「服装」「髪型」「化粧」「１０代」などは一般的な語と考えられるが，「ファッション」を検索語から除くと検索課題に適合しないと思われるため，「ティーンエイジャー」「X 世代」「ファッション」の 3 つの語を残し，再び検索実験を行った．その結果 Average precision の値は 0.0061 と低下した．そこで検索課題と正解の記事を見比べ，もう一度検討し直すと，動向が知りたいという課題なので，それに変わる「流行」という語を追加し 4 つの語で検索を行った．その結果，0.0878 と少し精度が向上した．

(3) 前述の 8 題において，不要と思われる一般的な語を検索語から削除し，検索実験を行った結果 Average precision の値は 0.3380 となり，検索精度が向上した．

4. 考察

検索語の改良により各検索課題において様々な変化が見られた．改良には一定の基準を設けてはいないが，例にある「逆エルニーニョの影響」のように，検索課題には記されていないが，別名「ラニーニャ」を検索語にすると，大幅に精度が向上するものや「ティーンエイジャーのファッション動向」のように広範囲で曖昧なものについては，検索語を改良してもあまり向上しないものがあった．精度低下の原因として考えられるのが，一般的な語の使用である．一般的な語で検索実験を行うと，正解記事以外も検索されるため，適合度が低下し，精度自体が低下する．しかし検索課題によっては，一般的な語を検索語とした場合でも精度が向上するものもあり，検索課題によっては，その語を必要としている場合があるため，一般的な語の扱う場合は注意する必要がある．

また，検索語を拡張した場合より，ある程度絞り込むことで精度が向上しているように感じる．以上の事から，検索精度を向上させるには，語の適正の他に，数などにも関わりがあると考えられる．

5. まとめ

今回の検索実験では，NTCIR-6 の結果からいくつかの検索課題を抽出し，それらの検索精度を向上させるため，検索語の改良とその評価を行った．その結果，調査した 8 題において検索精度の向上が見られた．

また現在では，この一連の動作をすべて手作業で行っているが，今後はシステムの自動化が必要になる．その場合，検索語は辞書ＤＢや情報サイトなどから検索課題に適した語を選択すればよいが，選択した語の適正評価や語の数を最適化する処理も必要となってくる．

システムを自動化する場合，このような評価や処理を一般化することは難しいが，ある程度の絞り込みを行い，その検索課題において，適切な語を選択するようにしなければならない．そのためには，システムの自動化方法を模索し，その方法に沿った検索実験を続けていく必要がある．

参考文献

[1] NTCIR HOME：http://research.nii.ac.jp/ntcir/index-ja.html