自然言語処理 121－20
情報学基礎 47－12
（1997. 9. 12）

情報フィルタリングシステムNEATのための検索要求文からのプロファイル生成

酒井哲也　梶浦正浩　住田一男

(株) 東芝 研究開発センター

我々は、新聞社・雑誌社により日々提供される電子化記事から個々のユーザーの興味に合ったものを選出し電子メイルなどで配信する情報フィルタリングシステムNEATを開発した。NEATは、プロファイルに記述されたブール式、検索語の出現位置・文書内密度・文書内分布などの多様な検索条件ベクトルに基づき、文書に対して加点しランキングを行う。今回、BMIR-J1の自然言語で書かれた検索要求文からプロファイルを自動生成する実験を行い、単純なブール式のプロファイルと人手によるプロファイルの中間程度の性能を達成できることを確認した。初期プロファイルの自動生成とrelevance feedbackの併用により、人手によるプロファイル作成の負荷は大幅に軽減されると考えられる。

Profile Generation from Query Sentences for the NEAT Information Filtering System

Tetsuya SAKAI　Masahiro KAJIURA　Kazuo SUMITA

Research and Development Center, Toshiba Corporation
1 Komukai Toshiba-cho, Saiwai-ku, Kawasaki 210, Japan

The NEAT information filtering system selects relevant articles from digital text provided daily by Japanese newspaper companies and publishers, and sends them by e-mail to its users. NEAT calculates a score for each article and produces a ranked output based on various types of query vectors written in the profile, such as location, density and distribution of keywords as well as boolean operators. We show that profiles generated automatically from query sentences can lie halfway between simple boolean profiles and hand-made profiles with respect to retrieval effectiveness. By combining this method and relevance feedback, the burden of manual profile definition will be lightened considerably.

1 まえがき

近年、テキスト情報の洪水の中からユーザーの欲しい情報のみを抽出する情報フィルタリングの技術が注目を集めている。我々は、新聞社・雑誌社により日々提供される電子化記事から個々のユーザーの興味に合ったものを選出し電子メイルなどで配信する情報フィルタリングサービスを実現するシステムとしてNEAT [1][1] を開発し、1996年4月に運用を開始した。NEATは、プロファイルに記述されたブール式、検索語の出現位置・文書内密度・文書内分布などの多様な検索条件ベクトルに基づき、文書に対して加点しランキングを行う。このように多様な観点から総合的に文書に対して加点を行う検索システムとしては、英語を対象とした [2] などが報告されているのみで、日本語を対象とした実運用システムは他に報告されていない。現在、NEATのプロファイルは人手により経験的に作成されている。そこで本稿では、プロファイル作成の負荷を軽減するために、自然言語で書かれた検索要求文から初期プロファイルを自動生成する試みについて報告する。

2 情報フィルタリングシステムNEAT

2.1 プロファイルで利用可能な検索条件

以下に、NEATのプロファイルに用いられる代表的な検索条件の種類について簡単に説明する。

bool条件 通常のAND,OR,NOTによるブール式。この検索条件が利用された場合、このブール式を満足した文書のみが他の検索条件による類似度計算の対象となる。

head/line(1)/para(1)条件 文書の見出し・第一文目・第一段落目に特定の語が出現した場合に得点を与える。

text条件 文書中に特定の語が出現した場合に得点を与える。ベクトル空間モデルによる点数付けに相当する。

ldens/cdens条件 特定の語の、文書中における行単位・文字単位の密度に応じて得点を与える。文書内で語が占める割合が高いほど高得点がつく。

[1] a News Extractor with Accurately Tailored profiles

ldist/cdist条件 特定の語の、文書中における行単位・文字単位の分布に応じて部分点を与える。語が均一に分布しているほど高得点がつく。

```
# DVD
bool,          DVD;
text:1,        DVD;
head:1,        DVD:2, プレーヤー:1;
ldens-text:2,  DVD;
text:-1,       DVD−ROM;
```

図1: プロファイルの例

図1にNEATのプロファイルの例を示す。各条件の直後に指定されている数値はプロファイル内でのその条件の重みであり、また各単語の直後に指定されている数値はその条件ベクトル内でのその単語の重みである(省略時は1)。

2.2 類似度計算方法

プロファイルと各文書との類似度の計算は、プロファイル中にbool条件がある場合にはこれを満たした記事のみに対して、ない場合にプロファイル中の語をひとつ以上含む記事に対して行われる。NEATはプロファイルと文書dとの類似度を以下のように計算する。

$$sim_c(i,d) = \frac{v(i)\cdot v_c(i,d)}{|v(i)||v_c(i,d)|} \quad (1)$$

$$sim_m(i,d) = \frac{v(i)\cdot v_m(i,d)}{|v(i)||v_m(i,d)|} \quad (2)$$

$$sim(d) = \frac{\sum_i w_c(i) sim_c(i,d) + \sum_i w_m(i) sim_m(i,d)}{\sum_i |w_c(i)| + \sum_i |w_m(i)|} \quad (3)$$

$v(i)$:プロファイル中のi番目の条件ベクトル
$v_c(i,d)$:条件ベクトルiに対応する文書dの頻度ベクトル(文字マッチ)
$v_m(i,d)$:条件ベクトルiに対応する文書dの頻度ベクトル(形態素マッチ)
$w_c(i), w_m(i)$:条件ベクトルiの重み(通常$w_c(i) = w_m(i)$)

3 検索要求文からのプロファイル生成

前節で示したように、NEATのプロファイルには人手によるきめ細かな記述が可能であり、正解記事を利用したチューニングに役立っている。しかし、未知のデータに対して初めて検索を行う場合、単純なプロファイルでもよいから自動生成できることが望ましい。そこで本稿では、自然言語で記述された検索要求文からプロファイルを自動生成する実験を行った。NEATは多様な検索条件に対応しているが、まずbool条件により類似度計算の対象となる記事を絞り込み、次にベクトル空間法[3]に近いtext条件を用い、さらにその他の条件により加点を行うというおおまかな方針を予め定め、主に以下の点に着目しながら段階的に実験を行い、有効なプロファイルの構成方法を探ることにした。

- 検索語として利用するのは、形態素解析結果がよいか、文字種切り[2]結果がよいか
- ひとつのtext条件に複数の検索語を並べて多次元ベクトルとしたほうがよいか、検索語をひとつだけ記述したtext条件を複数羅列するほうがよいか[3]
- bool+text条件と併用すると有効なのはどの検索条件か

プロファイルの有効性の指標としては、再現率-適合率曲線[5]を用いた。

4 評価用テストセット

4.1 BMIR-J1

検索システム評価用ベンチマークBMIR-J1[4][4]は、日本経済新聞経済面の600記事と60の検索要求と正解記事集合[5]により構成される。検索要求は難易度別に表1のようにグループ化されており、かなり難しい検索要求が多い。本稿ではこれら60件全てについて検索実験を行い、各再現率-適合率曲線を平均化した。

表1: BMIR-J1の検索要求グループと件数

グループ	要求される検索機能	件数
グループa	基本機能のみ	10件
グループb	数値・レンジ機能必要	5件
グループc	構文解析機能中心	6件
グループd	言語知識利用中心	12件
グループe	世界知識利用中心	10件
グループf	言語知識・知識処理併用	17件
		計60件

4.2 日本経済新聞CD-ROM1995年版

BMIR-J1は検索対象が600記事と少ないので、これとは別に日本経済新聞CD-ROM 1995年版の7月1日から10日までの約5,000記事に対しても同じ検索要求で検索実験を行った。ただし、正解記事の選定は全数調査ではなく、ブール検索にヒットした記事にのみついてBMIR-J1の正解判定基準にならって行った。よって、ここで得られた各正解集合は真の正解集合の部分集合であると考えられるが、再現率計算の際にはこれらを真の正解集合とみなした。なお、ブール検索は以下のように行ったものである。

1. 検索要求文の文字種切りおよび形態素解析を行う。
2. 上記結果の和集合から2文字以上の文字列を取り出し、ORで連結する。
3. NEATのシソーラス展開機能により各検索語を展開してからbool条件による検索を行う。

ただし、60件の検索要求のうち4件については上記の方法により正解を見つけることができなかったので、本稿ではこれらを除く56件を用いた。

[2] 漢字・ひらがな・カタカナ・英数字・記号などの文字種の変わり目を利用して単語切りを行うワンパスの単純かつ高速な処理。形態素解析と比較すると、辞書が不要・未知語に強い・語を長めに切り出すなどの特徴がある。

[3] このように構成されたプロファイルをそれぞれ多次元プロファイル、多条件プロファイルと呼ぶことにする

[4] 株式会社日本経済新聞の協力によって、社団法人情報処理学会・データベースシステム研究会・情報検索システム評価用データベース構築ワーキンググループが、1993年9月1日から12月31日の日本経済新聞記事を基に構築した情報検索評価用データベース（テスト版）。

[5] 正解レベルにはA,Bの2段階が用意されているが、本稿ではこれらを区別せず正解として扱った。

5 実験

5.1 bool条件による検索

bool条件は各種条件による文書のランキングの前処理として文書をある程度絞り込むために用いる。この段階では文書を多めに拾っておいて、最終的にランキング結果の上位のみを取るようにすればよいので、ここでは検索要求文から抽出した語をORで連結することにする。検索語としては文字種切り結果(表2の1)および形態素解析結果(同2)を採用した[6]。結果を図2および図5に示す。全体的に文字種切り結果のほうが性能がよいので、以後の実験ではbool条件には文字種切り結果を採用した。

5.2 bool+text条件による検索

bool条件で絞り込んだ文書に対しては、まずtext条件で加点することにする。ここでは、検索語として形態素解析結果か文字種切り結果のどちらを採用するか、また多次元プロファイルと多条件プロファイルのどちらにするかに応じて、表2の1-1〜1-4の4通りの実験を行った。結果を図3および図6に示す。両テストセットにおいて、形態素解析結果の多条件プロファイルが最も有効であることがわかる。

5.3 bool+text+もう一種の条件による検索

bool+text 条件に加えてもう一種類の条件を用い、性能向上を図る。text 条件以外の条件の場合も文字種切り結果を用いたり多次元ベクトルを構成したりすることは可能であるが、今回は text 条件にならい形態素解析結果の多条件プロファイルを構成することにした。head/line(1)/para(1)/ldens/cdens/ldist/cdist 条件をbool+text条件に加えて用いた場合について実験を行った(表2の1-1-1〜1-1-7)。結果を図4および図7に示す。語の密度により加点するldens/cdens条件が両テストセットにおいて有効であった。

[6] ここでは、長さが2以上の語のみを用いた。

表2: 各曲線の説明

記号	説明
1	bool(文字種切り結果のOR連結)
2	bool(形態素解析結果のOR連結)
1-1	1+text(形態素解析結果; 多条件)
1-2	1+text(形態素解析結果; 多次元)
1-3	1+text(文字種切り結果; 多条件)
1-4	1+text(文字種切り結果; 多次元)
1-1-1	1-1+ldens(形態素解析結果; 多条件)
1-1-2	1-1+cdens(形態素解析結果; 多条件)
1-1-3	1-1+para(1)(形態素解析結果; 多条件)
1-1-4	1-1+cdist(形態素解析結果; 多条件)
1-1-5	1-1+line(1)(形態素解析結果; 多条件)
1-1-6	1-1+ldist(形態素解析結果; 多条件)
1-1-7	1-1+head(形態素解析結果; 多条件)
h	人手で作成したもの
1A	1のグループaのみの平均
1-1-1A	1-1-1のグループaのみの平均
hA	hのグループaのみの平均

5.4 その他の組合せ・条件重みの変更

bool+text+ldens条件あるいはbool+text+cdens条件にさらに条件を加えたり、text 条件とldens/cdens条件の条件重みを変たりしてみたが、目だった性能向上は見られなかった。よって、検索要求文からプロファイルを自動生成する場合は、例えばbool+text+ldens条件を条件重みを全て1にして用いればよいと考えられる。

6 考察

図8および9に、検索要求文から自動生成したbool+text+ldens条件のプロファイルの性能をbool条件のみのプロファイルおよび検索要求文とコメントを読み人手で作成したプロファイル[6][7](表2のh)と比較したグラフを示す。

[7] 筆者が経験に基づき作成したもの。検索要求文に含まれない検索語も含んでいる。またbool条件を用いていないプロファイルもある。

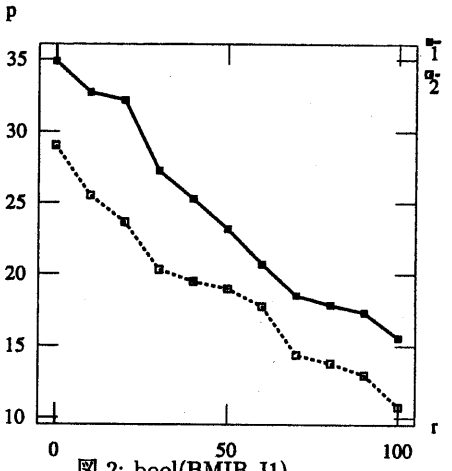

図 2: bool(BMIR-J1)

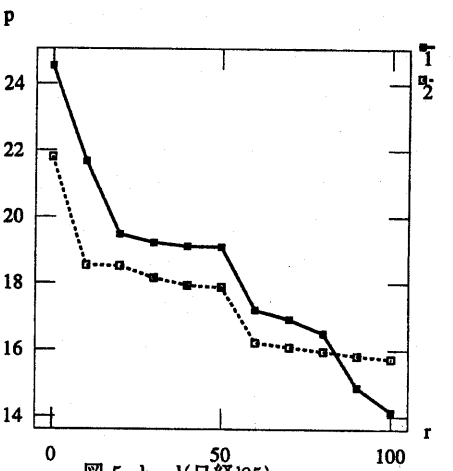

図 5: bool(日経'95)

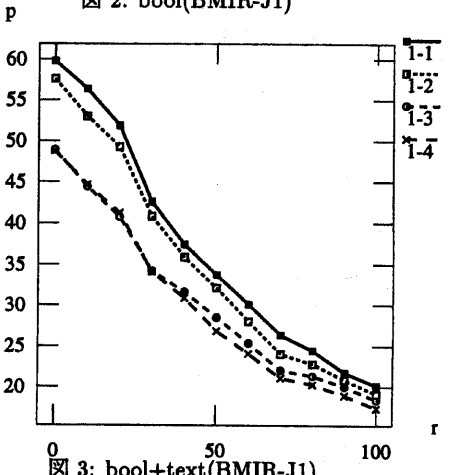

図 3: bool+text(BMIR-J1)

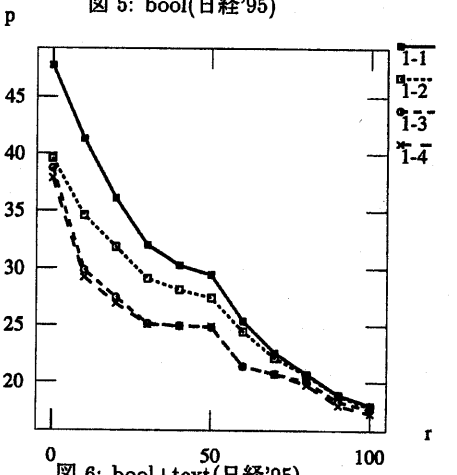

図 6: bool+text(日経'95)

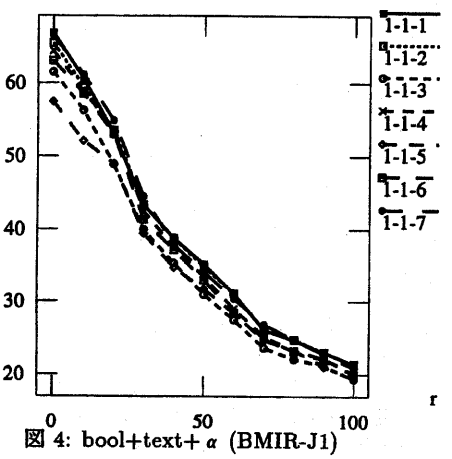

図 4: bool+text+ α (BMIR-J1)

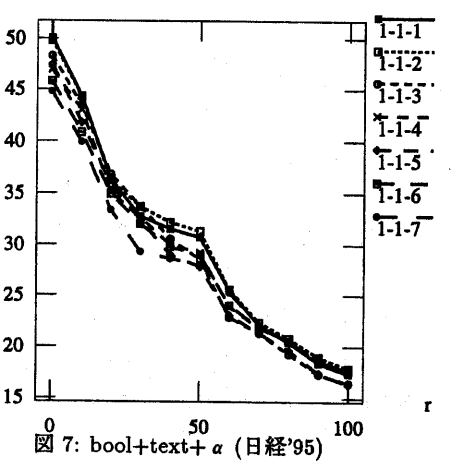

図 7: bool+text+ α (日経'95)

bool条件のみのプロファイルはいわば検索性能の下限であり、一方人手によるプロファイルはNEATの検索能力を経験的に最大限に活かしたという意味で上限といえる。自動生成したプロファイルにより両者の中間程度の性能が得られることがわかる。人手によるプロファイルは、検索要求文そのものだけでなく正解判定基準を詳細に記述したBMIR-J1のコメント情報も活用して作られたものであることもあり、自動生成したプロファイルもこの性能には及ばない。しかし、自動生成したプロファイルを初期プロファイルとして位置づけ、以後正解記事から新たな検索語を抽出してプロファイルのチューニングを行うようにすれば、人手によるプロファイル作成の負荷は大幅に軽減されると考えられる。

また、図8および9には、BMIR-J1の検索要求のうち検索語の有無の判定やシソーラス展開など基本的な機能を要求した検索要求グループaの10件(表1)のみについての比較結果も示している(表2のhA,1-1-1A,1A)。グループaの正解記事は、boolだけで比較的簡単に検索できてしまうため、text条件およびldens条件の貢献度が低い。

7 まとめ

情報フィルタリングシステムNEATのプロファイルをBMIR-J1の自然言語による検索要求文から自動生成する試みを行い、以下の知見を得た。

- bool+text+ldens/cdens条件の併用が有効
- 検索語としては、bool条件では文字種切り結果が、text条件では形態素解析結果が有効
- 多次元よりも多条件プロファイルのほうが有効
- 上記方法により自動生成したプロファイルは、bool条件のみのプロファイルと人手で作成したプロファイルの中間程度の性能を示す

ユーザーに負荷を与えない程度の、少ないrelevance判定結果をもとに、上記の方法で作成した初期プロファイルに対してrelevance feedback[3]を効率よく行う方法を開発することが今後の課題である。

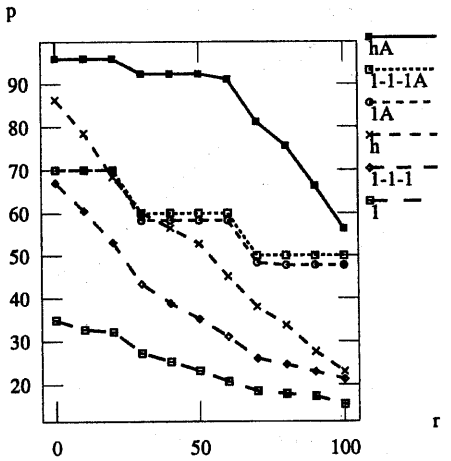

図 8: boolおよび人手との比較 (BMIR-J1)

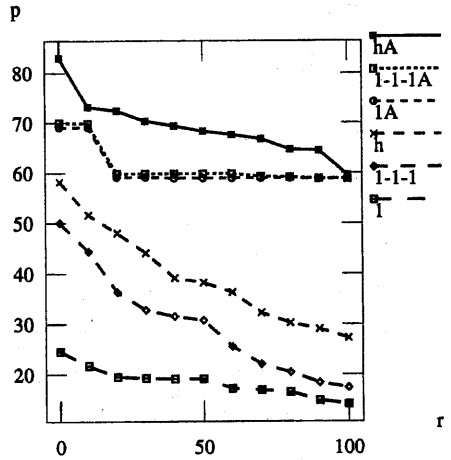

図 9: boolおよび人手との比較 (日経'95)

参考文献

[1] 梶浦他：“情報フィルタリングシステムNEATの開発,” 第54回情処学会全国大会, pp.3-299–300, 1997.

[2] J.P. Callan et al.: “The INQUERY Retrieval System,” In *Proceedings of the Third International Conference on Database and Expert Systems Applications*, pp.78–83, 1992.

[3] G. Salton ed.: “The SMART Retrieval System(邦訳: SMART情報検索システム),” Prentice-Hall,1971.

[4] 福島他：“日本語情報検索システム評価用テストコレクション BMIR-J1,” 自然言語処理シンポジウム, 1996.

[5] I.H. Witten et al.: “Managing Gigabytes: Compressing and Indexing Documents and Images,” Van Nostrand Reinhold, pp.148–151, 1994.

[6] 酒井他：“ベンチマーク BMIR-J1 を用いた情報フィルタリングシステム NEAT の評価,” 第54回情処学会全国大会, pp.3-301–302, 1997.