情報フィルタリングシステムNEATの開発

1S－10

梶浦正浩† 三池誠司‡ 酒井哲也† 佐藤誠† 住田一男†
†(株) 東芝 研究開発センター ‡(株) 東芝 Advanced-I 事業推進企画室

1. はじめに

我々は、新聞社/雑誌社などから日々提供される文書（記事）よりユーザの要求に合致するものを抽出しユーザに提供する、実サービス用の情報フィルタリングシステム **NEAT**(News Extractor with Accurately Tailored profiles) およびシステムの中核であるフィルタリングエンジンを開発した。フィルタリングエンジンは、2種類の単語検索方法を結合した新しい検索法や多様なフィールドに対応した複数の検索条件ベクトルを用いることによって、高い再現率/適合率を実現できるよう設計されている。本稿では、開発した **NEAT** およびフィルタリングエンジンの概要について述べ、また、新しい単語検索法の評価結果を示す。

2. NEATの概要

図1に **NEAT** の処理の概要を掲げる。**NEAT** は大きく分けて、24時間稼働する「受信-前処理部」と、一定時刻に起動される「フィルタリング-配信部」の2つに分けられる。前者は終日インターネット/ISDN回線/アナログ回線などを用いて記事プロバイダ（新聞社/雑誌社等）から送られてくる記事を受信し、受信ファイルを1記事毎に分割後、記事毎に形態素解析などの前処理および文字列検索インデックスの作成を行う。後者は、作成された記事DBやインデックスおよび、SQL Serverに格納されている個人情報や検索要求（プロファイル）を元にフィルタリングを行い、同一内容の記事の削除等を行った後、配信媒体に応じた記事の整形を行ってユーザに配信する。

3. フィルタリングエンジンの計算モデル

3.1 拡張点

我々が開発したエンジンのフィルタリングモデルは、一般的な単一ベクトルモデルではなく、以下のいくつかの点において拡張されている。

文字レベル検索/単語レベル検索

プロファイルに記述された単語とのマッチにおいて、文字単位での検索（文字レベル検索）は洩れがない代わりに過剰にマッチしてしまうことがある（例：「アマコスト駐日大使」に「コスト」がマッチ）。逆に形態素解析の結果と単語とを対照することによる検索（単語レベル検索）では先の例のような誤ったマッチが生じない代わりに形態素解析自体が失敗する可能性もある（例：「東京/都」と「東/京都」）。つまり、文字レベル検索と単語レベル検索は一長一短であり互いに補完する関係にある。そこで、本モデルでは両方の検索における類似度を荷重平均し文書の類似度の計算を行う方法を用いている。

複数の検索ベクトル

一般に文書の内容を表す単語は文書に一様に分布しているわけではなく、例えば最初の一文や一段落目、概要セクションなどに集中していることが多い。単一ベクトルの検索モデルやフィルタリングモデルでは、単語の出現位置の非一様性の取り扱いが難しい。そこで、本モデルでは文書内の各種フィールドに対応した検索ベクトルを設定することを可能にした。表1は開発したフィルタリングエンジンが取り扱うことのできるフィールドの種類である。

表 1: フィールドの種類

フィールドタイプ	フィールド
text	文書全体
abst	概要
intr	導入
head	見出し
sh	小見出し
ssh	小々見出し
sssh	小々々見出し
allh	全ての見出しおよび小見出し
allsh	全ての小見出し
line(N)	N文目
para(N)	N段落目

3.2 フィルタリングの流れ

プロファイルには例2のように、フィールド名及び単語を記述する。各単語には単語荷重 r_{ij} を付与する（省略時は1）。また、i 番目のフィールドには文字レベル検索及び単語レベル検索でのフィールド荷重 $W^c(i)$, $W^w(i)$ を付与する。例では $W^c(i)(=W^w(i))$ はそれぞれ 2, 3, 1, 1, 1 であり、検索荷重ベクトル $\boldsymbol{R}(i)(=(r_{ij}))$ はそれぞれ (1), (1), (3, 1), (2, 1, 1), (1, 1, 1) である。

文書 d の類似度 S_d は以下の式のように計算する。

$$S_d = \frac{\sum_i W^c(i)P_d^c(i) + \sum_i W^w(i)P_d^w(i)}{\sum_i |W^c(i)| + \sum_i |W^w(i)|} \quad (1)$$

Development of Information Filtering System **NEAT**

Masahiro KAJIURA, Seiji MIIKE, Tetsuya SAKAI, Makoto SATO, Kazuo SUMITA

Research & Development Center, Toshiba Corporation
1 Komukai-Toshiba-cho, Saiwai-ku, Kawasaki, 210, Japan
Tel: 044(549)2240, Fax: 044(520)1308
E-mail: kajiura@eel.rdc.toshiba.co.jp

$P_d^c(i)$, $P_d^w(i)$ は、

$$P_d^x(i) = \frac{\boldsymbol{R}(i) \cdot \boldsymbol{F}_d^x(i)}{|\boldsymbol{R}(i)| \cdot |\boldsymbol{F}_d^x(i)|} \quad (2)$$

で求める。ここで、$\boldsymbol{F}_d^c(i)$, $\boldsymbol{F}_d^w(i)$ は文字/単語レベル検索での単語の出現頻度ベクトルである。

4. 実験

本節では、文字レベル検索と単語レベル検索の結合がどれだけの効果が生むかを評価する。

4.1 実験方法

まず、(1) 式を以下のように変形する。

$$S'_d = \frac{\sum_i W^c(i)P_d^c(i) + \alpha \sum_i W^w(i)P_d^w(i)}{\sum_i |W^c(i)| + \alpha \sum_i |W^w(i)|} \quad (3)$$

α を 0 にした場合、類似度 S'_d は文字レベル検索のみの場合等価になり、∞ にした場合は、S'_d は単語レベル検索のみの場合と等価になる。

本実験では、1995 年 6 月および 7 月に発行された日本経済新聞 32481 記事および 302 のトピックに関するプロファイルを用い、α を 0, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0, ∞ に変化させ、各々における正規化適合率及び正規化再現率の平均を求めた。

4.2 実験結果

図 3に求まった正規化適合率及び正規化再現率の平均を掲げる。図には表示されていないが、$\alpha = \infty$ での正規化適合率/再現率は各々0.9916, 0.999959 であった。図より、$\alpha = 1$ つまり、文字レベル検索及び単語レベル検索の類似度を 1 対 1 で合成した類似度を用いることによって、より高い精度でのフィルタリングが可能であることがわかる。

5 まとめ

情報フィルタリングシステム **NEAT** および情報フィルタリングエンジンを開発した。フィルタリングエンジンは、(1) 文字レベル検索及び単語レベル検索の両方を用いる、(2) 文書の各種フィールドにおける類似度を用いることによって、フィルタリング精度が向上するよう設計されたモデルに基づき開発された。また、実験によりモデルで用いている単語検索方式により、フィルタリング精度が向上していることを確認した。なお、**NEAT** は 1996 年 4 月より実運用中である。

参考文献

[1] 酒井ほか, 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, 本大会予稿 1S-11, 1997.

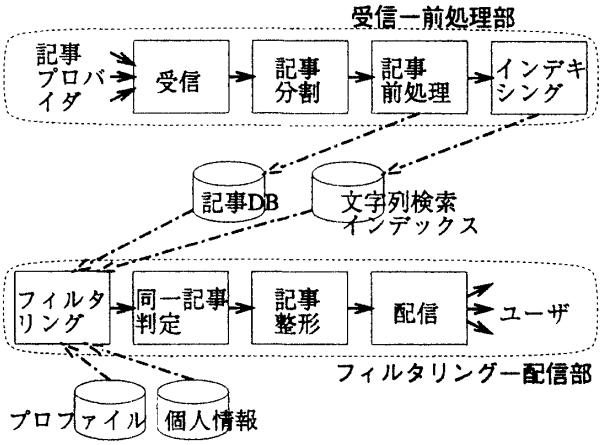

図 1: **NEAT** の処理の流れ

```
# レーザーディスク
text:2, レーザーディスク;
head:3, レーザーディスク;
body:1, レーザーディスク:3, プレーヤ:1;
body:1, レーザーディスク:2, 音楽, 映像;
body:1, レーザーディスク, レンタル, カラオケ;
```

図 2: プロファイルの例

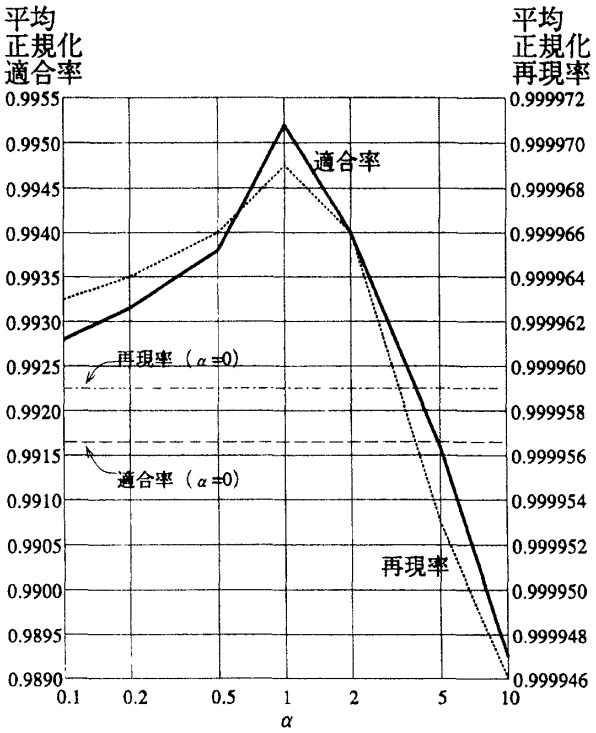

図 3: 実験結果