情 報 学 基 礎 43－13
（1996. 9. 13）

WWW上の電子新聞に対する情報フィルタリングとその評価

菅井 猛　和田 光教

沖電気工業株式会社　研究開発本部　マルチメディア研究所

〒108 東京都 港区 芝浦 4－10－3

Email: sugai@okilab.oki.co.jp, mwada@okilab.oki.co.jp

インターネットをはじめとする情報ネットワークで提供される情報源は大規模化しており、ユーザにとって役にたつ情報のみを取り出すことが重要になってきている。その一つの試みとして、WWW上の電子新聞に対して情報フィルタリングシステムを開発したので、そのアーキテクチャーについて述べる。また、情報フィルタリングの精度の評価方法を考察して、情報検索システム評価用ベンチマーク (BMIR-J1) を用いて、3つの関連フィードバックを評価した。さらに、日本語の情報フィルタリングにおいて文字の切り出し方法が精度に関連してくるが、新しい切り出し方法で基底語を決めてフィルタリングした評価結果について報告する。

キーワード
情報検索、情報フィルタリング

Information Filtering for Electronic Newspapers on the World-Wide Web and its Evaluation

Takeshi Sugai　Mitsunori Wada

Oki Electric Industry Co., Ltd.　Research & Development Group
Multimedia Laboratories

4-10-3, Shibaura 4-Chome, Minato-ku, Tokyo 108, Japan

Email: sugai@okilab.oki.co.jp, mwada@okilab.oki.co.jp

We have developed the information filtering system for electronic newspapers on the World-Wide Web and this paper describes the initial design of our information filtering system. We also studied the evaluation method of information filtering and evaluated the three relevance feedback by using information retrieval system benchmark(BMIR-J1). Because the Japanese extracting method is important for the effectiveness of information filtering, we developed the new extracting method. The paper shows the comparison of the experiment result for various extracting methods.

keyword
information retrieval, information filtering

1 はじめに

インターネットの普及により、ユーザは大量のテキスト文書を扱うことと、いかにユーザ自身にとって重要な情報のみを取得するかといったことが重要になってきている。このような情報氾濫を解決する1つのアプローチとして、情報フィルタリングの研究が近年盛んに行なわれている [1, 8, 9]。情報フィルタリングは、動的に追加されるテキストからユーザの興味に適合するテキストを抽出する技術である。その抽出のために、ユーザの興味を記述するプロファイルとテキストとの比較を行なう。その比較を行なうのに、情報検索の1つのモデルであるベクトル空間モデル [14] を用いる方法がある。ベクトル空間モデルでは、テキスト内に出現する語句に基づき特徴ベクトルを決定する。プロファイルとテキストの類似度を、対応する特徴ベクトルの類似度とみなして、ユーザの興味に近いテキストのみを抽出する。英語のテキストの場合、単語間が分かちがきされており、不要語の削除と stemming により、特徴ベクトルが計算される [20]。この時、日本語のテキストの場合、特徴ベクトルを計算するためには、形態素解析などの語句切り出し方法が必要である。特徴ベクトルをどのように決めるかは検索精度に密接に関係してくる。

本論文では、まず、我々が開発したネットワーク情報フィルタリングのアーキテクチャについて述べる。次に、情報フィルタリングの評価方法を考察し、Salton が提案している3つの関連フィードバック [16] について、WWW 上の電子新聞を用いた評価と、情報検索システム評価用ベンチマーク (BMIR-J1) を用いた評価について述べる。さらに、検索精度を向上するために、形態素解析で得られた基底語を、収集されたテキストの統計情報を利用することによって、形態素を統合して基底語を得る方法について提案する。この方法を、情報検索システム評価用ベンチマーク (BMIR-J1)[6] を用いて評価した結果について述べる。

2 ネットワーク情報フィルタリング

2.1 アーキテクチャ

ネットワーク情報フィルタリングシステムは、アクセス制御文によって指定された URL からテキストを収集して、ユーザのプロファイルに合わせてフィルタリングを行なう。ユーザのプロファイルを

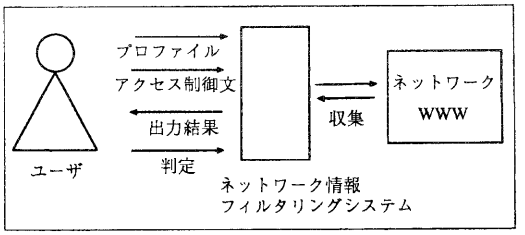

図 1: ネットワーク情報フィルタリングの入出力

満たした情報資源をユーザが評価することによって、プロファイルを書き換えて、テキストの収集とフィルタリングを行なう (図 1)。

ここで、プロファイルとは、ユーザの欲する用件を自然言語で表現したものである。また、アクセス制御文は、以下のパラメータからなる。

- 収集の始点
- 収集の深さ
- 間隔時間
- タイムアウトの時間
- 収集の戦略

収集の始点とは、情報を収集する開始点を示す URL である。

収集の深さとは、URL の始点から、収集するために辿るリンクの最大の距離である。ただし、再帰的なリンクは含まない。

間隔時間とは、収集を開始してから、再び情報を収集するまでの時間である。これは、WWW 上で定期的に変更される情報資源を収集する時に、有効である。例えば、WWW 上の電子新聞を一日一回収集するためには、24 時間と入力することによって、24 時間ごとに情報資源の収集を行なえる。

タイムアウトの時間とは、情報収集を開始してから情報収集を打ち切るまでの時間である。

収集の戦略とは、収集の方式を指定する。収集の方法は、「全経路収集」、「ヒューリステックスを用いた収集」がある。全経路収集は、深さの範囲内で、情報資源をすべて収集する方法である。ヒューリスティックスを用いた収集とは、深さの範囲内で、プロファイルを満たす情報資源の中のリンクをたどり収集を行なう方法である。

出力結果は、「プロファイルを満たしたテキスト」と「類似度」との組である。例えば、図 2のように HTML で記述したテキストに出力される。

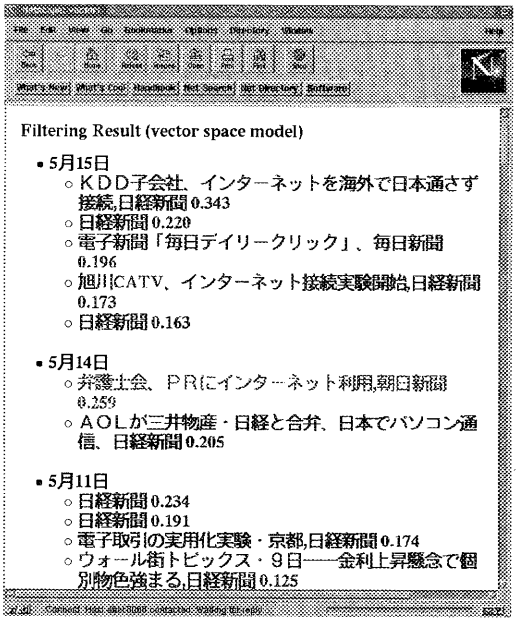

図 2: フィルタリング結果

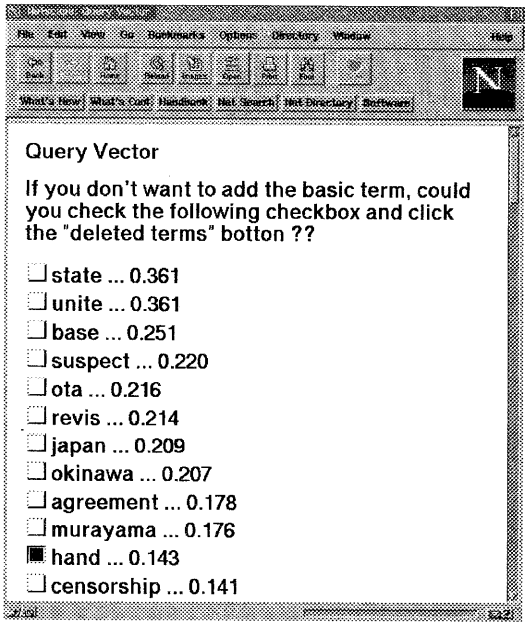

図 3: 基底語選定インタフェース

関連フィードバックにより、ユーザのプロファイルを書き換えるメカニズムだけでは、プロファイルにユーザの興味に関係ない基底語が入ってしまうことがある。そのため、ユーザが明示的に基底語を変更するためのインタフェースが必要になる (例えば、[5, 12])。図 3のインタフェースにより、ユーザは、不要である基底語を消去することができる。

2.2 フィルタリングエンジン

2.2.1 基底語の選定方法

英語のテキスト

テキストが英語の場合、基底語を求めるために、HTML で記述されたテキストから HTML のタグを取り除き、基底語にはなりえない不要語を削除する。この不要語の削除は、Fox の論文に掲載された 425 語の stoplists を用いた [4]。

不要語が削除された中から、stemming の処理を行なう。stemming とは、英語の活用形の語幹を基本形にする操作である。例えば、ties → tie, hopping → hop のように stemming を行なう。stemming のアルゴリズムは、Porter の stemming のアルゴリズム [11] を用いる。なお、Porter のアルゴリズムは、多くの stemming のアルゴリズムの中で、効率がよいことが知られている [20]。

日本語のテキスト

基底語の選定は、ベクトル空間モデルにとって、重要な役割を果たす。特に、日本語の場合、複合語に対して、基底語をどのようにとるかがフィルタリングの性能に大きく影響してくる。

試作システムでは、日本語の基底語の選定方法として、文字種の違いによって基底語を選定する方法と、形態素解析を利用した基底語の選定する方法について実装している。

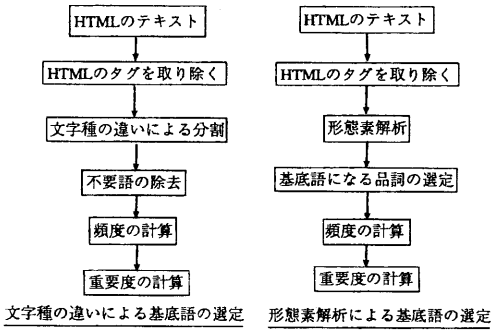

図 4: 日本における基底語の処理の流れ

文字種の違いによる基底語の選定

平仮名、カタカナ、漢字、英字、数字、その他の記号などの、文字種の違いにより基底語を決める。例えば、「米大統領選の序盤最大のヤマ場となるニューハンプシャー州予備選が、２０日に行われる。」という文について文字種の違いによって分割すると、以下のようになる。

米大統領選／の／序盤最大／の／ヤマ／場／となる／ニューハンプシャー／州予備選／が／、／２０／日／に／行／われる／。

さらに、文字種によって分割したものから、不要語を削除する。例えば、次の１文字は、ユーザにとって重要ではないというヒューリステックスによって基底語から削除する。

の、に、や、と、も、を、は、が

文字種の違いにより基底語を決めた場合は、なんらかのヒューリステックスを導入して精度を向上させる必要がある [18]。この処理をテキストのすべての文に適用して、そのテキストの全体の基底語の頻度を求める (図 4)。

形態素解析による基底語の選定

形態素解析ツールである JUMAN[7] を用いて形態素解析し、重要度を計算する。試作システムでは、名詞である形態素を基底語として用いている。

2.2.2 ベクトル空間モデル

ベクトル空間モデルは、テキストや質問文に出現する語句に基づき基底語を定め、テキストや質問文を基底語の張る空間内のベクトルとして特徴付け、その類似度により検索を行なう検索モデルである [16]。テキストに出現する各基底語に対して出現頻度を基にした重要度を算定し、テキストを表現するベクトルを決定する。質問文は自然言語文で表現されており、同様に重要度の算定によりベクトルを決定する。検索結果は、質問文に対して、類似度が大きいテキストの順にランキングされる。類似度は、テキストのベクトルとプロファイルのベクトルの内積で計算され、０から１までの値をとる。

基底語とは、通常、英語では単語である。一方、日本語では文字種の違いによって分割された文字列、あるいは、テキストを形態素解析して得られた単語である。

あるテキストのベクトルを次式で表現する。ここで、$dw_1, dw_2, ..., dw_t$ は、そのテキストの基底語の重要度である。

$$DW = (dw_1, dw_2, ..., dw_t)$$

同様に、質問文のベクトルを次式で表現する。ここで、$q_1, q_2, ..., q_t$ は、質問文の基底語の重要度である。

$$Q = (q_1, q_2, ..., q_t)$$

また、基底語の重要度 (テキスト D_i の語 T_k のベクトルの重み) は、以下の式で与えている。

$$W_{ik} = \frac{(tf_{ik}) \cdot (log\frac{N}{n_k})}{\sqrt{\sum_{j=1}^{t}(tf_{ij})^2 \cdot (log\frac{N}{n_j})^2}} \qquad (1)$$

ここで、各記号の意味は次の通りである。

- W_{ik} ... テキスト D_i の語 T_k のベクトルの重み
- tf_{ik} ... テキスト D_i の語 T_k の出現数
- N ... 収集されたテキストの総数
- n_k ... 収集されたテキストの中で、語 T_k が含まれているテキストの数

(1) 式の $log\frac{N}{n_k}$ は、一般に idf(inverse document frequency) と呼ばれる。情報フィルタリングにおいて、idf は、収集されたテキストの総数 (コレクションと呼ぶ) をどのようにとるかにより考慮に入れなくてはならない。Sheth のシステムでは、ネットニュースのフィルタリングの場合、流れている記事の特徴量があまり変化がないので、一週間に１回程度、idf を計算し直している [17]。また、Salton は、コレクションが頻繁に変更される場合は、このパラメータを無視することを推薦している [15]。

類似度は、テキストのベクトルとプロファイルのテキストの内積で計算され、以下の式で定義される。

$$Sim(DW, Q) = \sum_{i=1}^{t} dw_i \cdot q_i \qquad (2)$$

ここで、以下の θ のようなスレッシュホールドが存在する。検索結果は、式 (3) を満たしてものの中で、類似度が大きい順序に表示される。

$$Sim(DW, Q) > \theta \qquad (3)$$

2.2.3 関連フィードバック

試作システムでは、時間のともに変化するユーザの興味に追従するために、関連フィードバックを用いている。評価では、文献[16]で用いられている以下の関連フィードバックを使って評価を行なった。

Ide dec-hi 法

$$Q_{new} = Q_{old} + \sum_{all\_relevant} DW_i - \sum_{one\_nonrelevant} DW_i \quad \ldots (4)$$

ここで、*all_relevant* とは、検索された文書の中で、ユーザにとって興味がある文書すべての特徴ベクトルである。*one_nonrelevant* とは、検索された文書の中で、ユーザにとって興味がない文書の中で一番頻度の多い基底語のベクトルである。

Ide regular 法

$$Q_{new} = Q_{old} + \sum_{all\_relevant} DW_i - \sum_{all\_nonrelevant} DW_i \quad \ldots (5)$$

Standard Rocchio 法

$$Q_{new} = Q_{old} + \beta \sum_{n_1\_rel\_docs} \frac{DW_i}{n_1} - \gamma \sum_{n_2\_nonrel\_docs} \frac{DW_i}{n_2} \quad \ldots (6)$$

ここで、n_1*_rel_docs* とは、検索された文書の中で、ユーザにとって興味がある文書すべての特徴ベクトルである。n_2*_nonrel_docs* とは、検索された文書の中で、ユーザにとって興味がない文書すべての特徴ベクトルである。また、n_1, n_2 は、それぞれ、ユーザにとって興味がある文書の数、ユーザにとって興味がない文書の数である。また、β, γ は、0から1までの実数であり、$\beta + \gamma = 1.0$ である。試作システムでは、Salton ら[16]が用いたパラメータ $\beta = 0.75, \gamma = 0.25$ として評価した。

3 評価

情報フィルタリングの評価方法として、本稿では、検索精度の評価を行なった。情報フィルタリングでは、「時間とともに変化する、ユーザの興味にいかに追従するか」、「ある時間において、ユーザのプロファイルを満たすようにフィルタリングされているか」という2つを考えなければならない。後者について、情報検索における再現率(recall)、適合率(precision)の評価方法を適用することにより評価を行なった。

3.1 WWW 上の電子新聞に対する評価

3.1.1 評価方法

1996年1月に、毎日1回、WWW上の電子新聞を収集して、フィルタリングを行なった。関連フィードバックの方法、式(4)、(5)、(6)において、日本語の場合はどのような結果がでるかについて評価した。フィードバックによる繰り返しを1、2、3回、行なった時の検索精度を比較した。

WWW上の電子新聞では、一つのテキストの中に複数の記事がある場合、テキストの中の記事を分割して、各記事ごとに新たなテキストを作成してそれぞれのテキストごとにフィルタリングを行なった。

3.1.2 評価

表1、表2にそれぞれ、再現率と適合率を示す。プロフィル文は、「村山政権 政権 自民党 連立政権」である。式(3)のスレッシュホールドの値として0.1を与えた。

表 1: 再現率 (Recall) の比較

		初期	1回	2回	3回
Ide dec-hi	(1)	0.333	0.333	0.833	0.750
Ide dec-hi	(2)	0.778	0.600	0.833	0.500
Ide regular	(1)	0.333	0.333	0.600	0.250
Ide regular	(2)	0.778	0.600	0.833	1.000
Rocchio	(1)	0.333	0.333	0.667	0.500
Rocchio	(2)	0.778	0.600	0.833	1.000

(1)... 文字種の違い, (2)... 形態素解析

3.1.3 考察

- 試作システムでは、フィードバックを何回もかけると、元のプロファイルの語の重要度が際だたなくなり、適合率が落ちることが確認された。なお、Salton によると、関連フィードバックでは、2、3回フィードバックをかけた時が一番よい結果がでている[3, 13]。
- 文字種の違いによる基底語の選定では、形態素解析による基底語の選定と比べて、検索精度は劣る。

表 2: 適合率 (Precision) の比較

		初期	1回	2回	3回
Ide dec-hi	(1)	0.750	0.333	0.556	0.167
Ide dec-hi	(2)	1.000	0.200	1.000	0.333
Ide regular	(1)	0.750	0.333	0.571	0.143
Ide regular	(2)	1.000	0.200	1.000	0.364
Rocchio	(1)	0.750	0.333	0.667	0.182
Rocchio	(2)	1.000	0.200	1.000	0.364

(1)... 文字種の違い, (2)... 形態素解析

- 文字種の違いによる基底語の選定の場合、上記のヒューリステックスだけだと条件が弱過ぎる。例えば、「い」、「き」、「しい」、「え」、「である」、「な」、「り」などが基底語になってしまう。
- Ide dec-hi 法と Ide regular 法の違いは、nonrelevant な文書の基底語をどの程度フィードバックに反映するかである。この実験では、nonrelevant な文書が比較的少なかったので、データ上の違いはそれほど見られなかった。

3.2 評価用ベンチマークに対する評価

評価用ベンチマークに対する評価として、情報検索システム評価用ベンチマーク Ver.1.0 (BMIR-J1)[1]を用いて評価を行なった。BMIR-J1 は、新聞記事 600 件、検索要求文 60 件、および各検索要求文に対する正解集合から構成されている。正解集合は、基準として、正解、主題も一致 (ランクA)、正解、主題は別 (ランクB)、不正解、全く関係ない記事 (ランクC) のように 3 つのランクに分かれている。本評価では、ランクA、ランクBを正解とする。また、本評価では、BMIR-J1 の中で、表層的な質問文のみを評価の対象とする。

3.2.1 評価方法

BMIR-J1 の記事を HTML で記述されたテキストに変換して評価を行なった (図 5)。

[1]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) を利用

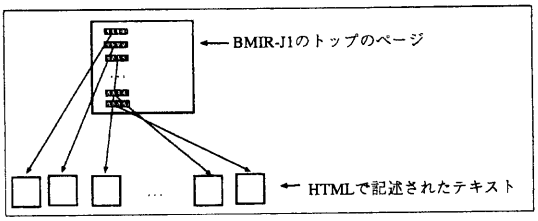

図 5: 評価データの構造

情報フィルタリングではある時間におけるユーザのプロファイルを満たすための検索精度は、基底語の選び方 (品詞の選び方、複合語の分解方法) や基底語の重要度の計算方法に依存する。このような条件を同じにして、関連フィードバックのどの方法 [16] が日本語のデータに対して有効なのかを評価した。

3.2.2 複合語の扱い

日本語において、複合語をどのように分割するかは、検索精度に密接に関係してくる [10]。大量のデータから抽出した統計量を利用して、複合語を決める様々な方法が提案されている [19]。我々は、小川の方法 [10] を形態素レベルに適用することにより、形態素を統合する方法に対して実験を行なった。

この方法では、「2 形態素の間を統合する確率 (統合確率と呼ぶ) は、前の形態素の形態素尾確率と後ろの形態素の形態素頭確率の積である」ことを仮定する。ここで、形態素の形態素頭 (尾) 確率とは、基底語になりえる連続した形態素の先頭 (末尾) になる確率である。これは、その形態素が先頭 (末尾) である連続した形態素の総出現頻度を、収集されたテキストにおけるその形態素の総出現頻度で割った値として計算される。

テキストを形態素解析して基底語の候補を選定した後、隣り合った基底語を統合するかを上記の方法で決める。例えば、「3 期以上連続の減益企業」という質問文の場合、形態素解析により、以下のように分割される。

3 期 以上 連続 nil 減益 企業

ここで、 nil は、その前の基底語とその後ろの基底語が統合することがないことを示す。そして、隣り合った基底語になりえる形態素の統合確率は以下のようになる。

3 0.032 期 0.055 以上 0.136 連続 nil
減益 0.450 企業

統合確率が処理パラメータとして用意されているスレッシュホールドより小さいものを統合する。この時、スレッシュホールドとして 0.08 を与えれば、上記の例では、以下のような基底語が得られる。

3期以上、 連続、 減益、 企業

形態素の総出現頻度は、idf と同じように、追加されるテキストの一連した流れから取得される。上記の方法を形態素統合方式と呼ぶ。

3.2.3 評価

最初の実験では、関連フィードバックをかける前で、最初のプロファイル時における、文字種の違いによる方法、形態素解析によって得られた名詞を基底語として使う方法、形態素統合方式について検索精度を調べた (図 6)。なお、形態素統合方式のスレッシュホールドを 0.08 に設定した。ここでは、BMIR-J1 の質問文の中から、表層的な検索機能だけで検索できる質問文でかつ、2 個以上の基底語からなる質問文を 10 個を選んで、再現率、適合率の関係を求めた。

図 6のグラフによると、形態素解析によって得られた名詞を基底語として使う方法が一番検索精度がよいことがわかる。

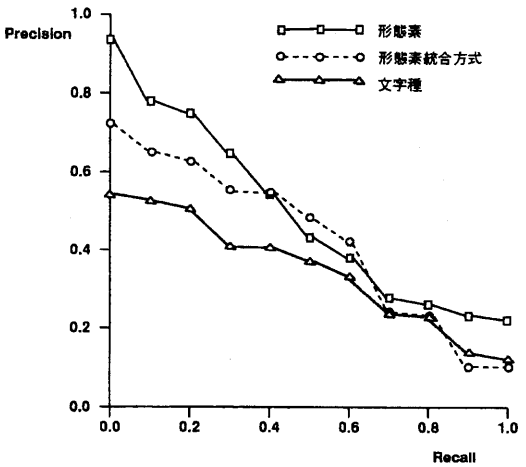

図 6: 文字切り出し方法の評価

2番目の実験では、形態素解析によって得られた名詞を基底語として使う方法で、関連フィードバック (4)、(5)、(6) の違いを比較した (図 7)。この結果によると、Rocchio 法が比較的よい結果を示している。

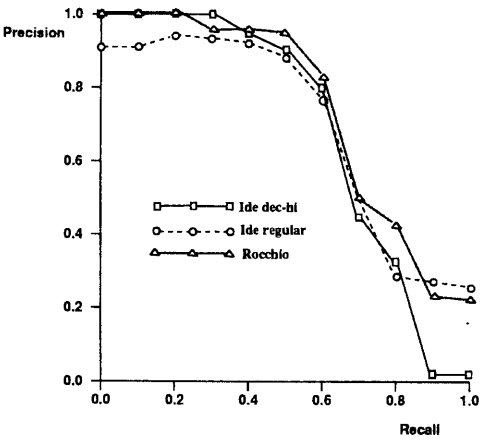

図 7: 関連フィードバックの評価

3番目の実験は、Rocchio 法 (6) 式における、文字種の違いによる方法、形態素解析によって得られた名詞を基底語として使う方法、形態素統合方式について検索精度を調べた (図 8)。 ここで、上位 10 までにランキングされたテキストの中から、BMIR-J1 の正解であるテキストをユーザにとって興味があるものとみなし、不正解であるテキストをユーザにとって興味がないものとみなして、フィードバックをかけた。また、それぞれのテキストごとに、重要度の大きい基底語、上位 10 個をフィードバックに用いた。

この実験では、形態素統合方式は検索精度はよくなかったが、今後、大規模なデータから統計的なデータを取って実験を行ない有効性を検討していきたい。

4 まとめ

本稿では、我々が提案するネットワーク情報フィルタリングのアーキテクチャについて述べた。また、情報フィルタリングにおいて、日本語の文字の切り出し方法を考察し、日本語における関連フィードバックについて評価を行なった。今後は、似たような概念を持つ基底語をまとめて一つのベクトルの軸にすることのできる、LSI(Latent Semantic Indexing) 法 [2] を、日本語の情報フィルタリングに適用して評価していく予定である。

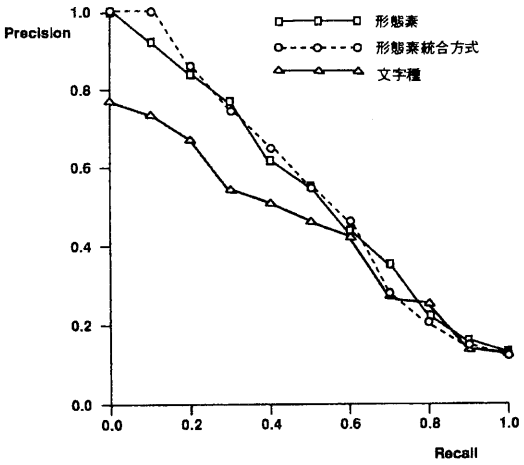

図 8: Rocchio 法の評価

参考文献

[1] Nicholas J. Belkin and W. Bruce Croft. Information Filtering and Information retrieval: Two Sides of the Same Coin? *Communication of the ACM*, Vol. 35, No. 12, pp. 29–38, December 1992.

[2] S. Deerwester, S. T. Dumais, G. W. Fornus, T. K. G. W. Landauer, and R. Harshman. Indexing by Latent Semantic Analysis. *Journal of the Society for Information Science*, Vol. 41, No. 6, pp. 391–407, 1990.

[3] Christos Faloutsos and Douglas Oard. A Survey of Information Retrieval and Filtering Methods. Technical report, University of Maryland, 1995. http://www.enee.umd.edu//medlab/filter/filter_project.html.

[4] Christopher Fox. Lexical Analysis and Stoplists. In *Information Retrieval*, pp. 102–130. Prentice Hall, 1992.

[5] Koenemann J. and Belkin N. J. A Case For Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness. In *CHI'96 Proceedings*, 1996.

[6] 芥子育雄, 他. 情報検索システム評価用ベンチマーク Ver.1.0 (BMIR-J1) について. 情報処理学会研究報告, Vol. DB 106-19, pp. 139–145, 1996.

[7] 松本裕治, 他. 日本語形態素解析システム JUMAN 使用説明書 version2.0, 1993.

[8] Masahiro Morita and Yoichi Shinoda. Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval. In *SIGIR'94 Proceedings*, 1994.

[9] Douglas Oard and Gary Marchionini. A Conceptual Framework for Text Filtering. Technical Report EE-TR-96-25, CAR-TR-830, CLIS-TR-96-02, CS-TR-3643, University of Maryland, 1996. http://www.ee.umd.edu/medlab/filter/papers/filter.ps.

[10] 小川泰嗣. 文字成分表を用いた効率的文書ランキング法の提案. アドバンス・データベース・シンポジウム' 95, pp. 29–38, December 1996.

[11] M. F. Porter. An Algorithm for Suffix Stripping. *Journal of the Society for Information Science*, Vol. 14, No. 3, pp. 130–137, 1980.

[12] Sheldon M. S., A. Duda, R. Weiss, and K. Gifford. Discover: a resource discovery system based on content routing. *Computer Networks and ISDN Systems*, Vol. 27, pp. 953–972, 1995.

[13] G. Salton. Relevance feedback and the optimization of retrieval effectiveness. In *SMART Retieval System - Experiments in Automatic Document Processing*. Prentice Hall Inc, 1971.

[14] Gerard Salton. *Automatic Text Processing, - The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Company, 1989.

[15] Gerard Salton and C. Buckley. Term Weighting Approaches in Automatic Text Retrieval. Technical Report 87-881, Department of Computer Science, Cornell University, November 1987.

[16] Gerard Salton and Chris Buckley. Improving Retrieval Performance by Relevance Feedback. *Journal of The American Society for Information Science*, Vol. 41, No. 4, pp. 288–297, 1990.

[17] Beerup Dilip Sheth. A Learning Approach to Personalized Information Filtering. Technical report, Master Thesis, MIT, Febrauary 1994.

[18] 須藤真理, 横尾英俊. 情報検索とデータ圧縮とを統合したシステム mg の日本語化. 情報処理学会研究報告, Vol. 95-FI-40, pp. 33–40, 1995.

[19] 小林義行, 徳永健伸, 田中穂積. 名詞間の意味的共起情報を用いた複合名詞の解析. 自然言語処理, Vol. 3, No. 1, pp. 29–43, 1996.

[20] Frakes W. B. Stemming Algorithms. In *Information Retrieval*, pp. 131–160. Prentice Hall, 1992.