WWW上の電子新聞に対する情報フィルタリング†

4T－8

菅井 猛*　和田 光教◇　森田 幸伯*

沖電気工業株式会社　研究開発本部§

1　はじめに

インターネットに代表される情報ネットワーク時代の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。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。この時、特徴ベクトルを計算するためには、日本語のテキストの場合には、形態素解析などの語句切り出し方法が必要である。

本稿では、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、情報フィルタリングの評価方法を考察する。また、文字種の違いによって基底語を決める方法と、形態素解析によって基底語を決める方法について、WWW上の電子新聞を用いて評価した。さらに、Saltonが評価している3つの関連フィードバック[3]を、情報検索システム評価用ベンチマーク(BMIR-J1)[1]を用いて評価した。

2　ネットワーク情報フィルタリング

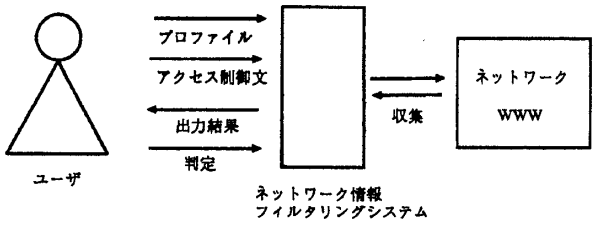

図1: ネットワーク情報フィルタリングの入出力

WWWのURL(Uniform Resource Locator)を始点として、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ネットワーク情報フィルタリングシステムを開発した。ユーザのプロファイルを満たしたテキスト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る。(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、フィルタリングする対象のテキストの起点を示すURLや収集する範囲を示したものである。

3　ベクトル空間モデルと関連フィードバック

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[3]。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。

試作システムでは、時間と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。試作システムでは、**Ide dec-hi 法**、**Ide regular 法**、**Standard Rocchio 法**という3種類の関連フィードバック[3]を実装している。

4　基底語の選定方法

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、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。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、形態素解析によって基底語を選定する方法について述べる。

文字種の違いによる方法では、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分割して基底語を求める。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、数字、記号などを不要語として削除する。

形態素解析による方法では、形態素解析ツールであるJUMAN[2]を用いた。形態素解析した品詞の情報を用いて重要度を計算する。試作システムでは、名詞である形態素を基底語としている。

---

†Information Filtering for Electronic Newspaper on the World-Wide Web

Takeshi Sugai*(E-mail sugai@okilab.oki.co.jp), Mitsunori Wada◇, Yukihiro Morita*

§Oki Electric Industry Co., Ltd., Research & Development Group

5 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。

5.1 WWW 上の電子新聞に対する評価

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。表1は、類似度がある値 (0.1) の時の再現率と適合率を示す。異なった3つのプロファイル文を対象とし、3種類の関連フィードバックの結果と2種類の切り出し方法を比較している。表1の各値は、3つのプロファイルに対する平均値である。プロファイル文は、2つ以上の基底語からなる。

文字種の違いによる方法は、形態素解析をした時に比べて、検索精度は劣る。また、3つの関連フィードバックの違いは、ユーザの興味に無関係なテキスト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、ユーザに無関係なテキスト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。

表 1: 再現率 (R)/ 適合率 (P) の比較

		R	R	P	P
		初期	1回	初期	1回
Ide dec-hi	(1)	0.230	0.328	0.917	0.694
Ide dec-hi	(2)	0.635	0.606	0.924	0.634
Ide regular	(1)	0.230	0.328	0.917	0.694
Ide regular	(2)	0.635	0.557	0.924	0.632
Rocchio	(1)	0.230	0.313	0.917	0.778
Rocchio	(2)	0.635	0.606	0.924	0.636

(1)... 文字種の情報のみ,(2)... 形態素解析したもの
「初期」は通常のフィルタリングを示し、「1回」は関連フィードバックを1回かけたことを示す

5.2 評価用ベンチマークに対する評価

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、複合語の扱い方) 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が日本語のデータに対して有効なのかを評価した。

BMIR-J1[1]のテキストを HTML で記述されたテキストに変換して、試作システム上でフィルタリングを行なうことによって評価を行なった。基底語は、JUMAN で解析された名詞を基底語とし、テキスト中のすべての基底語をベクトルとして登録する。BMIR-J1 の中から表層的な検索機能だけで検索できる検索文6個を選んで、再現率、適合率の関係を求めた (図2)。この実験では、Rocchio 法が比較的よい結果を示している。なお、Salton が英語のデータに対して行なった実験では、Ide dec-hi 法が優れている [3]。

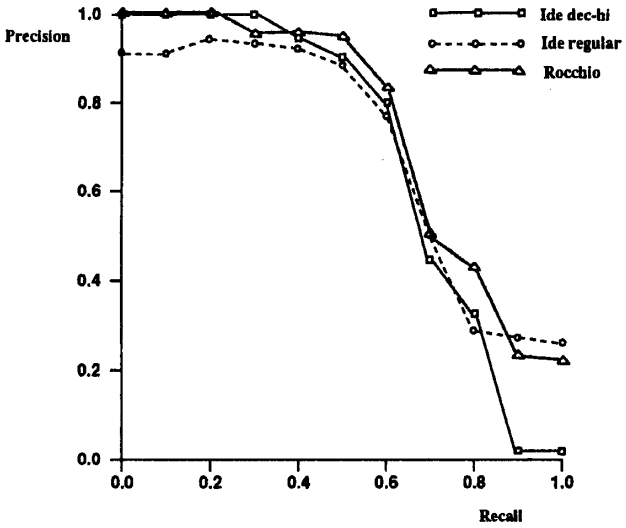

図 2: 関連フィードバックの評価

6 おわりに

本稿では、我々が提案するネットワーク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検討し、検索精度の評価を行なった。さらに、関連フィードバックを日本語のテキストを対象として、検索精度の評価を行なった。

参考文献

[1] 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.

[2] 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.

[3] 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.

[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