2008－NL－188 (2)
2008／11／26

相互類似性と配信時間差に基づくWebニュース冗長記事のフィルタリング

吉田 光範[†] 藤沢 匡哉[‡] 八嶋 弘幸[‡]

[†]東京理科大学 工学研究科 経営工学専攻

[‡]東京理科大学 工学部

近年，World Wide Webへのアクセス環境の多様化と普及により，Webによるニュース閲覧は一般化してきている．ニュースポータルサイトを使うことで，各々のWebニュースサイトに個別にアクセスすることなく，幅広い内容のニュースを一度に閲覧することができる．一方，複数のWebニュースサイトが配信するニュース記事の中には，しばしば同じ情報が記述された記事（冗長記事）が存在している．そのため，大量の冗長記事の中から，ユーザが自身の関心がある記事のみを選別するには手間がかかる．

本研究では，ニュース記事間の相互類似性と配信時間差に基づく2つの冗長記事フィルタ手法を提案する．提案手法は，配信時間差に基づき冗長性の判定に制限を加えることで，続報記事を選別除去することなく，冗長記事のみを選別除去する．実際のWebニュース記事を使った実験により，提案手法が冗長記事と続報記事を区別し，従来法よりも高い精度で冗長記事のみを選別できることを示す．

Redundancy news filtering based on similarity and published time difference.

Mitsunori Yoshida[†], Masaya Fujisawa[‡] and Hiroyuki Yashima[‡]

[†]Graduate School of Engineering, Tokyo University of Science

[‡]Faculty of Engineering, Tokyo University of Science

Recently, WWW (World Wide Web) is widely spread all over the world. It is getting common to browse news articles on WWW. In the news portal site, we can read news articles which plural news sites published, without visiting in those news sites. However, it is often seen that news articles with the same information appear in several news sites. Therefore, it is difficult to select only appropriate news articles with interest among large amount of redundant articles.

In this paper, we propose two redundancy news filtering methods which are based on similarity and published time difference. These filtering methods considers the published time of the news article in order to limit the article to judge as the redundant article. The first method uses window function and the second method uses machine learning. In the experiment, we show that our methods enable to distinguish redundant articles and follow-up articles.

1 はじめに

近年，World Wide Webへのアクセス環境の多様化と普及により Web によるニュース閲覧は一般化してきており，パソコンや携帯電話を使って，時と場所を選ばず最新ニュース閲覧が可能となっている．

ニュースポータルサイトは，新聞社などのWebニュースサイトが配信するニュース記事を集めて配信するサイトである．各々のWebニュースサイトに個別にアクセスすることなく，幅広い内容のニュースを一度に閲覧することができる．一方，複数のWebニュースサイトが配信するニュース記事の中には，しばしば同じ

情報が記述された記事（冗長記事）が存在している．ニュースポータルサイトではそれらが同時に配信されているため，これら大量の冗長記事の中から，ユーザが自身の関心がある記事のみを選別するには手間がかかる．

Web ニュース記事の選別に関して，我々はユーザのブックマークに登録されたページからユーザの関心情報を抽出し，それに合致するニュース記事を選別するフィルタ手法を提案した[5]．この手法は 1 つの Web ニュースサイトから配信されるニュース記事を対象として，ユーザの関心情報との類似性に基づき選別を行う．しかし，この手法をニュースポータルサイトに適用した場合には，フィルタ出力の中に冗長記事が含まれてしまうという課題が残されている．

冗長性と新規性の検出に関する研究に，過去の閲覧記事との類似性に基づき新規記事を選別する Zhang らの手法[1]がある．これは 2 段階の情報選別フィルタで構成され，第 1 段では情報ストリームの中から，ユーザの関心との類似性に基づき情報を選別し，続く第 2 段では，第 1 段の選別結果の中から記事間の類似性に基づき冗長性のある情報を選別除去する．しかし，この手法では，類似性のみに基づいて選別しているため，続報記事を考慮することはできないという問題がある．あるニュース話題に対して，最初に報じられる初報記事と，継続的に報じられる続報記事は，ニュース話題を追いかけるために必要な記事である．しかし続報記事と冗長記事は，共に初報記事と共通の話題を含むことから同程度の類似度があるため，記事間の類似度のみで判断するとこれらを区別できない．

本研究は，[1]における問題を解決し，ニュースポータルサイトで配信されている Web ニュース記事の中から冗長記事のみを選別除去する 2 つの冗長記事フィルタ手法を提案する．提案手法は，ニュース記事間の類似性に加えて配信時間差に基づき冗長性の判定に制限を加えることで，続報記事を除去することなく，冗長記事のみを選別除去する．

- 窓関数を用いたフィルタ（提案法 1）
- Support Vector Machine(SVM)を用いたフィルタ（提案法 2）

2 提案するフィルタ手法

2.1 フィルタシステムの概要

本論文で提案するフィルタシステムの概要を図 1 に示す．ある一つの社会的イベントに対し，複数の新聞社などにより配信されるニュース記事は，ニュースポータルサイトに集められ再び配信される．このとき冗長記事が含まれる．

図中の破線部分が本研究の提案システムであり，関心記事フィルタ（Interest filter)と冗長記事フィルタ（Redundancy filter）の 2 つの情報選別フィルタから構成する．はじめに，[5]で提案した関心記事フィルタにより，ユーザのブックマークに登録された Web ページとの類似性に基づき，ユーザの関心に合致する記事のみを選別する．この選別結果の中には，初報記事，冗長記事，続報記事が含まれるため，続く冗長記事フィルタにより冗長記事のみ選別除去する．最後に，初報記事と続報記事を関心記事（Interest）としてユーザに配信することで，フィルタリングが完了する．これ以降では，冗長記事フィルタで使用するフィルタ手法について述べる．

2.2 冗長記事フィルタ

一般に，ある初報記事に対する冗長記事は比較的短い時間差で配信され，初報記事と冗長記事の間の類似度は高い．それに比べ，続報記事はある程度，時間が経過してから配信され，初報記事と続報記事との類似度は中～高程度と考えられる．

図 2 に，冗長記事と続報記事の分布と，本研究およ

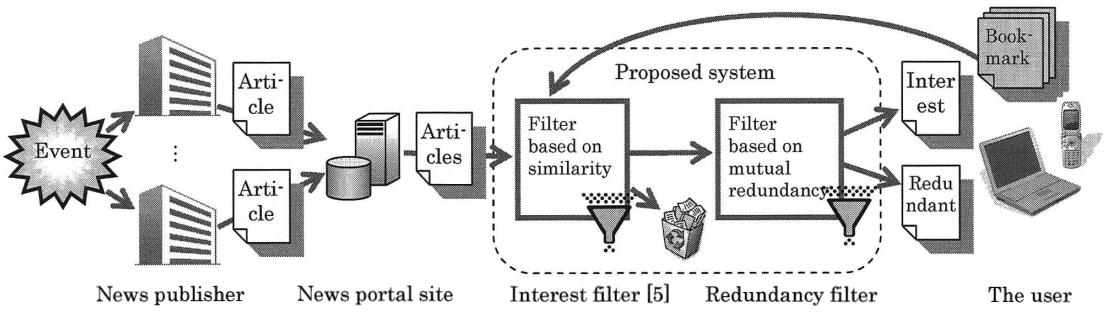

図 1: The outline of the proposed filtering system

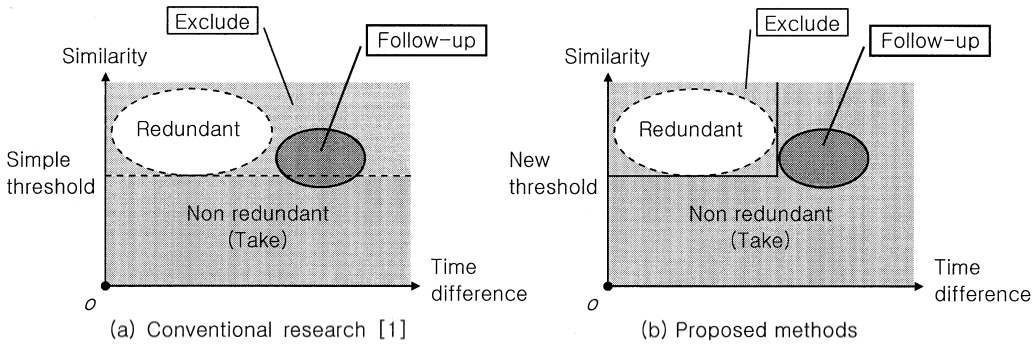

図2: Distribution of redundant articles and follow-up articles.

び Zhang らの手法[1]における冗長記事フィルタの概要を示す．縦軸が記事間の類似度，横軸が配信時間差であり，破線に囲まれた範囲が冗長記事の分布，実線に囲まれた範囲が続報記事の分布を表す．[1]では，類似度のみを尺度として閾値によって選別を行うため，続報記事を冗長記事として選別する誤りが生じる．一方，本研究の手法は，記事間の類似度に加えて記事の配信時間差を用いることで，続報記事を除去することなく，冗長記事のみを除去する．

続いて，冗長記事フィルタにおける類似度の算出方法を 2.3 節に，配信時間差に基づく窓関数を使って特定の配信時間差内のみを冗長記事の判定対象とする手法（提案法 1）を 2.4 節に，類似度と配信時間差の 2 次元分布に基づき機械学習を使って選別する手法（提案法 2）を 2.5 節に説明する．

2.3 ニュース記事間の類似度の算出

本研究では，文書間の類似性の尺度としてベクトル空間モデル（Vector Space Model）におけるコサイン距離を使用し，これを類似度と呼ぶ．

ニュース記事 D に対して，D に出現する単語 t_i を要素としてもつベクトルを単語ベクトル $\mathbf{t} = (t_i)$ と定義する．ここで i は単語番号とする．また，ニュース記事 D における t_i に対応する単語の出現頻度を要素としてもつベクトルを文書ベクトル $\mathbf{d} = (d_i)$ と定義する．これらのベクトルは，ニュース記事 D を形態素解析し，1 つの形態素を 1 つの単語として作成する．

ニュースポータルサイトで配信された Web ニュース記事から文書ベクトル $\mathbf{d}_j$ を作成し，n 個の Web ニュース記事群の文書ベクトル集合 $\mathbf{D} = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n\}$ を作成する．このとき j は文書番号とし，単語ベクトル $\mathbf{t}$ は文書ベクトル集合 $\mathbf{D}$ に含まれる全ての単語を要素とする．また，文書ベクトル集合 $\mathbf{D}$ に対応する重みベクトル集合を

$$\mathbf{G} = \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_n\} \tag{1}$$

とする．このとき $\mathbf{g}_j = (g_j^i)$ は重みベクトルと呼び，以下に示す tf 値，idf 値から定義される $g_j^i = \mathrm{tf}_j^i \times \mathrm{idf}_i$ を要素とする．

$$\mathrm{tf}_j^i = \frac{\log_2(d_j^i + 1)}{\log_2(\mathrm{tnum}(\mathbf{d}_j))} \tag{2}$$

$$\mathrm{idf}_i = \log_2 \frac{n}{\mathrm{dfreq}(t_i)} + 1 \tag{3}$$

ここで，各変数・関数は次のとおりである．

$\mathrm{tnum}(\mathbf{d}_j)$	文書 $\mathbf{d}_j$ における単語の異なり数
$\mathrm{dfreq}(t_i)$	$\mathbf{D}$ における単語 t_i を含む文書数
tf_j^i	文書 $\mathbf{d}_j$ における単語 t_i の出現頻度を正規化した値（単語頻度）
idf_i	$\mathbf{D}$ における単語 t_i の出現する文書の偏りの大きさ（文書頻度）

2 つのニュース記事の文書ベクトル $\mathbf{d}_a$ と $\mathbf{d}_b$ の間の類似度は，対応する 2 つの重みベクトルの間のコサイン距離

$$\mathrm{SIM}(\mathbf{d}_a, \mathbf{d}_b) = \frac{\mathbf{g}_a \cdot \mathbf{g}_b}{|\mathbf{g}_a||\mathbf{g}_b|} \tag{4}$$

と定義する．SIM は 0 から 1 の間の実数値をとり，SIM が大きいほど 2 つのニュース記事の類似性は大きいといえる．

2.4 窓関数を用いるフィルタ手法（提案法 1）

提案法 1 では，類似度 SIM を基に，特定の配信時間差内のみを冗長性の判定対象とするよう制限するために窓関数（Window Function）を使用する．窓関数は，ある区間以外で 0 となる関数である．

ニュースポータルサイトの配信記事を観察すると，複数の Web ニュースサイトから互いに同一の情報を含んだ冗長記事が短い時間差の間に配信されている．またその一方で，地方紙などの Web ニュースサイトでは，全国紙や通信社が配信したニュース記事が，比較的長い時間差を空けて再配信されることがある．従って，中程度の類似度を持つ冗長記事がしばしば短い時間差で配信され，高い類似度を持つ冗長記事がしばしば長い時間差の後に再配信される．このような構造に合わせた窓関数を選択するために，本研究では 3 つの窓関数（矩形窓 $\mathrm{WF_r}(t_{\mathrm{DIFF}})$，三角窓 $\mathrm{WF_t}(t_{\mathrm{DIFF}})$，ガウス窓 $\mathrm{WF_g}(t_{\mathrm{DIFF}})$）を使用し，実験によって現実の Web ニュース記事への当てはまりを評価する．窓関数はそれぞれ次のように表される．

$$\mathrm{WF_r}(t_{\mathrm{DIFF}}) = 1, \text{if } 0 \le t_{\mathrm{DIFF}} \le T_{\mathrm{RECT}} \tag{5}$$

$$\mathrm{WF_t}(t_{\mathrm{DIFF}}) = 1 - \frac{|t_{\mathrm{DIFF}}|}{T_{\mathrm{TRI}}}, \text{if } 0 \le t_{\mathrm{DIFF}} \le T_{\mathrm{TRI}} \tag{6}$$

$$\mathrm{WF_g}(t_{\mathrm{DIFF}}) = \exp\left(-\left(\frac{t_{\mathrm{DIFF}}^{\ 2}}{\sigma^2}\right)\right), \text{if } 0 \le t_{\mathrm{DIFF}} \tag{7}$$

ここで T_{RECT} は時間軸方向の窓の幅，T_{TRI} は三角窓における底辺の長さ，σ^2 はガウス分布の分散，t_{DIFF} は時間とする．図 3 に，横軸を時間 t_{DIFF}，縦軸を関数値としたときの，それぞれの窓関数の概形を示す．

フィルタリング処理をされる記事の文書ベクトル $\mathbf{v} \in \mathbf{D}$ が

$$[f_1(\mathbf{v}) = \max_{\mathbf{u}\in\mathbf{D},\mathbf{u}\neq\mathbf{v}} \{\mathrm{SIM}(\mathbf{u},\mathbf{v}) \times \mathrm{WF}_{q\in\{r,t,g\}}(\mathrm{TIMEDIFF}(\mathbf{u},\mathbf{v}))\}] \ge \theta$$

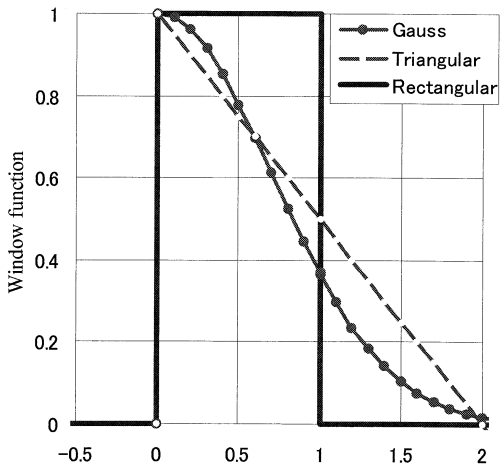

図 3:Shapes of window functions

のとき $\mathbf{v}$ を冗長記事に選別する．ここで $\mathrm{TIMEDIFF}(\mathbf{d}_a\mathbf{d}_b)$は文書ベクトル $\mathbf{d}_a$ と $\mathbf{d}_b$ に対応する 2 つのニュース 記事間の配信時間差とし，θ は閾値とする．

2.5 SVM を用いるフィルタ手法（提案法 2）

提案法 2 では，2 つのニュース記事間の類似度と配信時間差からなる 2 次元空間上のデータをサポートベクターマシン(SVM)を用いて学習および分類することで冗長記事を選別する．2 次元空間上のデータは，

$$\mathbf{x}_l = (\mathrm{TIMEDIFF}(\mathbf{d}_a,\mathbf{d}_b), \mathrm{SIM}(\mathbf{d}_a,\mathbf{d}_b)), \tag{8}$$

$$(l = 1,\ldots,\frac{n(n-1)}{2}),\ 1 \le a < b \le n,\ a,b \in \mathbf{Z}$$

とする．SVM は，データを識別超平面により 2 つのクラス（class）に分離する機械学習であり，学習時に与えるクラス y_l は，文書ベクトル $\mathbf{d}_a$ と $\mathbf{d}_b$ に対応する 2 つのニュース記事の関係により

$$y_l = \begin{cases} +1\text{:冗長記事} \\ -1\text{:それ以外} \end{cases} \tag{9}$$

とする．y_l は $\mathbf{x}_l$ に対応し，学習は式(8)，式(9)の対応するデータから L 個を抽出して行う．

2.2 節の図 2 に示したように，冗長記事の選別のためには非線形の分離が必要である．また，冗長記事と続報記事の分布が一部重なることにより，学習データにノイズが含まれることが考えられる．本研究では，ノイズを許容するために制約を緩めるソフトマージンと，非線形分離のための高次元写像を使用する．

識別誤りを表すスラック変数（slack variable）$\xi_l \ge 0 (l = 1,\ldots,L)$ と制約を緩和するパラメータ C を導入し，最適化問題を次のように設定する．目的関数は

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{l=1}^{L}\xi_l, \tag{10}$$

であり，制約条件は

$$y_l(\mathbf{w}^T \cdot \mathbf{x}_l + b) \ge 1 - \xi_l, (l = 1,\ldots,L). \tag{11}$$

となる．ここで $\mathbf{w}$ は超平面の直交する方向を定義する 2 次元ベクトルであり，b は超平面の原点からの距離を表す．また高次元空間への写像の際に使用するカーネル関数 $K(\mathbf{x},\mathbf{x}')$ には，次の RBF(Gaussian Radial Basis Function) カーネルを用いる．

$$K(\mathbf{x},\mathbf{x}') = \exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2). \tag{12}$$

ここで最適化問題にラグランジュ乗数 $\alpha(\ge 0)$ を導入すると，式(10)，(11)は以下の双対問題に帰着される．

目的関数は

$$\max \sum_{l=1}^{L} \alpha_l - \frac{1}{2} \sum_{l,k=1}^{L} \alpha_l \alpha_k y_l y_k \mathbf{x}_l \cdot \mathbf{x}_k , \quad (13)$$

であり，制約条件は

$$0 \le \alpha_l \le C, (l = 1,..., L), \sum_{l=1}^{L} \alpha_l y_l = 0 , \quad (14)$$

となる．カーネル $K(\mathbf{x}, \mathbf{x}')$ の導入は，目的関数（13）を，

$$\max \sum_{l=1}^{L} \alpha_l - \frac{1}{2} \sum_{l,k=1}^{L} \alpha_l \alpha_k y_l y_k K(\mathbf{x}_l, \mathbf{x}_k) , \quad (15)$$

に変更することで行う．

提案法 2 では，あらかじめ L 個の学習データにより上記を解いて α，b を求めておく．続いてフィルタリング処理をされる記事の文書ベクトル $\mathbf{v} \in \mathbf{D}$ が

$$\left[f_2(\mathbf{v}) = \max_{\mathbf{u} \in \mathbf{D}, \mathbf{u} \ne \mathbf{v}} \left(\sum_{l=1}^{L} \alpha_l y_l K(\mathbf{x}_l, (\mathrm{TIMEDIFF}(\mathbf{u}, \mathbf{v}), \mathrm{SIM}(\mathbf{u}, \mathbf{v}))) + b \right) \right] \ge 0$$

のとき $\mathbf{v}$ を冗長記事に選別する．

3 評価実験

3.1 実験概要

ニュースポータルサイトで配信されている実際の Web ニュース記事を使用し，提案手法のフィルタ性能を評価する．ある期間に配信された全ての Web ニュース記事の中から，あらかじめいくつかのニュース話題を抽出しておき，さらにその中から冗長記事，初報記事，続報記事をラベル付けする．冗長記事をクラス Redundant，その他の全ての記事をクラス Other とし，この 2 つを正解クラスとして実験に使用する．

実験は 10 分割クロスバリデーションで行う．すなわち，正解クラスの 9/10 のデータを使用して 3.2 節および 3.3 節で説明する方法でパラメータ設定と SVM の学習を行い，残りの 1/10 のデータを使って 2 つの提案法のフィルタ処理を行う．それぞれの処理結果と正解カテゴリを表 1 に示す方法で比較して，

表 1: Comparison of filtering results

		Filtering results	
		Without redundant	With redundant
Correct class	Other	f (True Negative)	α (False Positive)
	Redundant	β (False Negative)	t (True Positive)

$$recall = \frac{t}{t+\beta} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}},$$

$$precision = \frac{t}{t+\alpha} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}},$$

$$F = \frac{2 \times precision \times recall}{(precision + recall)},$$

を計算し，10 パターンの実験の平均値を評価値として用いる．ここで *recall* は正解クラスの再現性を表す再現率，*precision* はフィルタ結果中の正解率を表す適合率となる．*recall* と *precision* はトレードオフの関係となり，F 値は総合的なフィルタ性能を表す．

実験にはニュースポータルサイト ceek.jp news で配信されている日本語の Web ニュース記事を使用し，その他の前提条件は次に示す通りとする．

【前提条件】

(1) **D**: ceek.jpから配信されたニュース記事のうち，条件(2)のサイトから配信された21243件から作成（期間：2007年8月26日～9月8日）
(2) 対象Webニュースサイト：ceek.jpでの配信件数が上位15位のサイト
(3) 単語ベクトルの構成要素：名詞，動詞，形容詞，副詞（[6]を参考に選択）
(4) 正解クラスRedundant：97個のニュース話題に含まれる冗長記事386件
(5) 実験方法：10分割クロスバリデーション

3.2 窓関数パラメータの設定

はじめに，提案法 1 で使用するパラメータの設定を次の 2 つのステップにより行う．

Step1: 閾値 θ の設定

続報記事と冗長記事の分布が重なる範囲では閾値の決定が難しいため，それらの分布が重ならない範囲として配信時間差が 3 時間以内の記事を使って閾値 θ を設定する．2.4 節の矩形窓を T_{RECT}=3 に設定して使用し，θ を変化させたときの選別結果から最大の F 値をとる θ を設定する．

Step2: T_{RECT}，T_{TRI}，σ の設定

矩形窓における T_{RECT} と評価値の関係を図 4 に示す．F 値の最大値はピーク形状となっており，このとき総合的なフィルタ性能が最大となるので，T_{RECT} にはその点の値を使用する．続いて，三角窓における T_{TRI} と評価値の関係を図 5 に示す．F 値は最大値のピークが明確ではなく，平らな頂点をとる．そこで *recall* と *precision* のバランスを考慮し，min(|*recall*-*precision*|)と

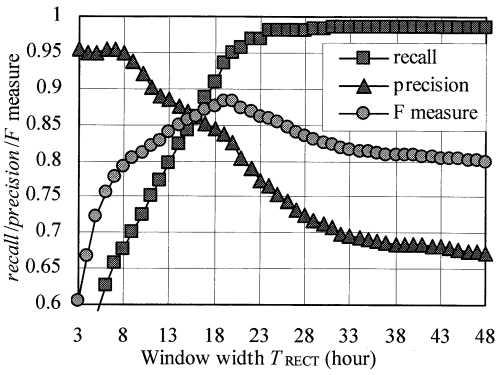

図 4: The relation of window width T_{RECT} and filtering performance of the method 1 with rectangular window function.

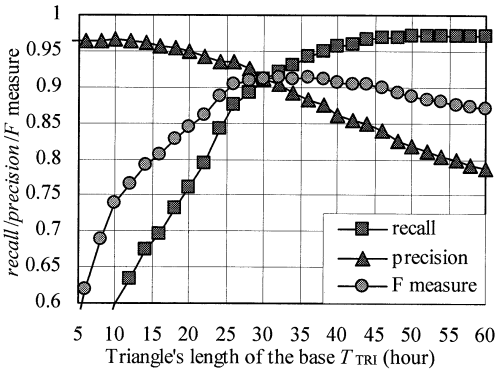

図 5: The relation of length T_{TRI} and filtering performance of the method 1 with triangular window function.

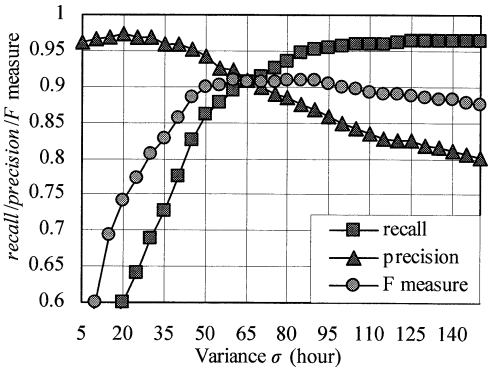

図 6: The relation of variance σ and filtering performance of the method 1 with gauss window function.

なる点の T_{TRI} を設定する．ガウス窓における σ と評価値の関係を図 6 に示す．ガウス窓は三角窓と同様に F 値が平らな頂点となったため，同様の手順により σ を設定する．

3.3 SVM パラメータの設定

続いて，提案法 2 で使用する RBF カーネルのパラメータ γ と，ソフトマージンのパラメータ C を設定する．γ={0.1,1,10,100,1000} と C={1,10,100,1000,10000}（[7]を参考に選択）の組み合わせにおける F 値を表 2 に示す．このとき最も F 値の高い組み合わせである γ=0.1，C=1000 を設定する．

4 実験結果

提案法 1，提案法 2，および従来法[1]における 10 分割クロスバリデーション平均値を表 3 に示す．

[1]は，F 値が 0.75 と最も低い．これは *precision* が 0.67 と低いためであり，続報記事を含む非冗長記事を，冗長記事として選別してしまう誤りが多いことを示す．

提案法 1 では，三角窓とガウス窓を使用したときに *precision* が 0.90 と最も大きな値となっている．大きな *precision* の値は，非冗長記事を冗長記事に選別してしまう誤りが少ないことを示す．また図 5，図 6 において F 値が平らな頂点をとっていることから，パラメータの変化に対する感度が低く，三角窓とガウス窓ではパラメータをある程度，正確に設定しておけば高い選別性能が得られることがわかる．一方，矩形窓は F 値が 0.87 と高いものの，図 4 において F 値の頂点がピーク形状を示しており，パラメータ T_{RECT} の変化に対する感度が高く，高い選別性能を得るためにはパラメータを正確に設定する必要がある．

提案法 2 は，総合的な評価値である F 値が 0.89 と最も大きな値となっている．また，*recall* (0.90) と *precision* (0.89) が共に高い値をとっていることから，SVM はニュース記事群の中から冗長記事のみを高い精度で選別できることを示している．

表 2: F measure of filtering performance with mutual combination of γ and C.

		γ				
		0.1	1	10	100	1000
C	1	78.69	84.20	86.22	77.43	56.37
	10	87.87	88.52	85.52	79.28	59.67
	100	88.50	88.12	83.91	78.91	60.01
	1000	**89.24**	87.41	83.60	77.68	59.88
	10000	89.06	86.79	82.55	78.27	59.83

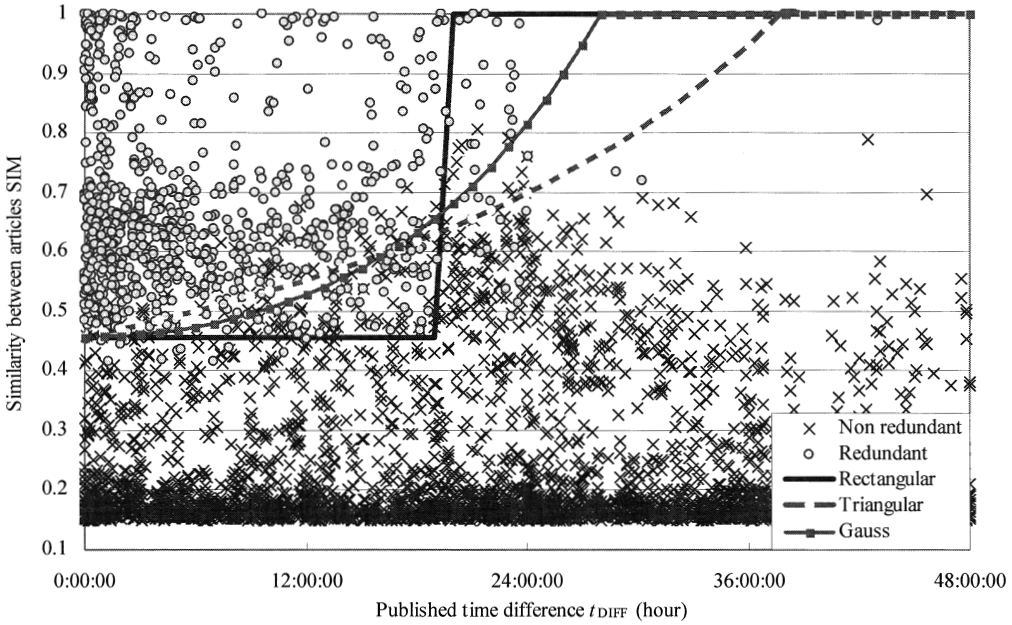

図 7: Relation of window functions to redundant articles distribution.

図 7 に，実験で使用した窓関数と，正解クラスの冗長記事分布の関係を示す．横軸が記事の配信時間差 t_{DIFF}，縦軸が類似度 SIM，○印が冗長記事，×印が非冗長記事であり，窓関数と SIM の積が閾値 θ を超える範囲として

$$\mathrm{SIM} \geq \frac{\theta}{\mathrm{WF}_{q\in\{r,t,g\}}(t_{\mathrm{DIFF}})}$$

が窓関数を使用した際の実際の選別閾値となる．図 7 において実線が矩形窓，破線が三角窓，点付線がガウス窓における選別閾値である．このうち三角窓とガウス窓が，冗長記事（Redundant）の分布に対してあてはまりの良いことが確認できる．一方，t_{DIFF} が 9〜18 時間，SIM が 0.5〜0.65 の範囲付近においては，冗長記事と非冗長記事（Non Redundant）の分布が重なっている．これらの分布が重ならないような類似度の計算方法を使用することができれば，提案法 1 の選別精度はさらに向上すると考えられる．そのような類似度を設定することが今後の課題である．

5 まとめ

本研究では，ニュースポータルサイトで配信されている Web ニュース記事の中から冗長記事のみを選別除去することを目的として，2 つの冗長記事フィルタ手法を提案した．提案手法は，記事間の相互類似性と

表 3: Average performance of different redundancy thresholds measured on Japanese real Web news data.

			Precision	*Recall*	*F*
Conventional method[1]			0.67	0.89	0.75
Proposed methods	Window function (method 1)	Rectangular	0.83	0.93	0.87
		Triangular	0.90	0.85	0.87
		Gauss	0.90	0.86	0.88
	SVM (method 2)		0.89	0.90	0.89

配信時間差に基づいて選別を行う．実際の Web ニュース記事を使った実験によって，第 1 の手法は，冗長記事の分布に対してあてはまりが良く，高い *precision* 値が得られた．第 2 の手法は，学習機械を使用することで，高い F 値が得られた．

以上のことから，相互類似性と配信時間差に基づくフィルタ手法により，従来法よりも高い精度で冗長記事のみを選別するフィルタの構築が可能となった．

参考文献

[1] Zhang, Y., Callan, J. and Minka, T. "Novelty and Redundancy Detection in Adaptive Filtering", Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.81-88, 2002.

[2] V. Lavrenko, J. Allan, E. DeGuzman, D. LaFlamme, V. Pollard and S. Thomas, "Relevance models for Topic Detection and Tracking", Proceedings of the Conference on Human Language Technology (HLT), 2002.

[3] 谷村正剛,田中(石井)久美子,中川裕志，”異なる発信元からの WWW ニュース記事の内容に基づく対応付け”，情処学研報, Vol.2001, No.112, 2001-NL-146, pp.89-94， Nov. 2001.

[4] Nello Cristianini and John Shawe-Taylor, “An Introduction to Support Vector Machines and other Kernel-based learning methods”, Cambridge University Press, 2000

[5] Mitsunori Yoshida, Masaya Fujisawa, and Hiroyuki Yashima，"RSS filtering based on bookmark information."，2007 RISP International Workshop on Nonlinear Circuits and Signal Processing，pp.417-420, May 2007.

[6] 松尾豊,石塚満，”語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム”，人工知能誌 Vol.17,No.3,pp.217-223，May 2002.

[7] 山名美智子,村田博士,小野田崇,大橋徹,加藤誠二，"腕金錆画像に基づく再利用判定精度の向上"，第 18 回人工知能学会全国大会(JSAI2004)，2F2-02, 2004.

[8] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, “Topic detection and tracking pilot study: Final report” In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[9] F. Sebastiani, “Machine learning in automated text categorization”, ACM Computing Surveys (CSUR), Vol. 34, Issue 1, Mar. 2002.

[10] 梶浦正浩,三池誠司，"ニューズウォッチ社における情報フィルタリングシステム"，情報管理, Vol.47, No.4, pp.267-274，2004.

[11] 向井誠,青野雅樹，“RSS に基づく個人向け内容型情報推薦プロトタイプシステム”，情処学研報, Vol.2005, No.94, pp.27-pp.32, Sept. 2005.

[12] 青野雅樹，"大規模文書からのアウトライヤー文書クラスター検出方法の研究"，電気通信普及財団研究調査報告書, No.20, pp.474-486，2005.

[13] 佐藤吉秀,川島晴美,佐々木努,奥雅博，"時系列ニュース記事における最新話題語抽出方法"，情処学研報, Vol.2005, No.73, pp.1-6, 2005-NL-168-(1), 2005.

[14] 森正輝,三浦孝夫,塩谷勇，"時制クラスタのトピック追跡"，第 17 回データ工学ワークショップ (DEWS2006), 6A-i5，2006.

[15] 大島裕明,小山聡,田中克己，"文書群を問合せとした兄弟カテゴリー文書の検索"，信学論(D), Vol.J90-D, No.2, pp.196-208，Feb. 2007.