E-002

ブックマーク情報に基づく RSS 情報のフィルタリング

A RSS filtering based on bookmark information.

吉田 光範† 藤沢 匡哉‡ 倉林 有‡ 八嶋 弘幸‡
Mitsunori Yoshida Masaya Fujisawa Yew Kurabayashi Hiroyuki Yashima

1. まえがき

1.1. 研究背景

近年、インターネットへのアクセス環境の多様化と普及により、パソコンや携帯電話の Web ブラウザを使い、時と場所を選ばず最新ニュース閲覧が可能となってきた。その一方、Web ニュースは更新量が多く"情報の洪水"という問題が生じてきている。例えば、新聞社の Web サイト asahi.com [*1] では約 100 件/日、ニュース・ポータル・サイト ceek.jp news [*2]では約 2,000 件/日のニュースが発信されている。ユーザはその中から自分の関心のあるニュース記事を自力で選別しなければならないが、多くの Web ニュース・サイトはニュース記事タイトルの時系列による一覧表示が行われているため記事の重要度やボリュームを把握しにくく、選別のためのユーザ負担は大きい。

こうした状況に対し、ニュースの選別を自動的に行うフィルタが実用化されている。代表的なものとして、特定カテゴリのニュースを抽出するセクションフィルタ、特定キーワードを含むニュースを抽出するキーワードフィルタが挙げられる。例えば google news[*3]では、この２つのフィルタを組み合わせて適用することで、各々のユーザに対しカスタマイズしたニュース一覧の提示を試みている。しかし効果不十分という問題点が指摘[1]されており、またキーワードを継続的にメンテナンスすることは困難である。

これらの問題点を考慮し、本研究は「ユーザが関心のあるニュース」を抽出する、以下の２つのフィルタ手法を提案・検証する。

(1) 新規ニュースの中から「ユーザの関心」に類似するニュースを抽出するフィルタ（提案法１）
(2) (1)に加え、さらに「ユーザの関心に関連する話題」のニュースを幅広く抽出するフィルタ（提案法２）

これらの手法により「ユーザが、負担なく自分に必要なニュースを読める」仕組みの構築が可能になる。

1.2. 先行研究

ユーザの関心に合致するニュース記事の推薦に関する研究の一つに、ユーザが過去に閲覧した全ての Web ページ群の内容をユーザが関心のある情報と考え、それと類似度が高い新規ニュース記事をユーザに提示する方法[2]がある。しかしこの方法ではユーザの関心の取得方法に課題が残る。ユーザは、必ずしも関心のある Web ページのみを閲覧しているのではない。例えば検索過程においては、内容が未知の Web ページを順次閲覧する行動をとる。従って閲覧履歴の中には、ユーザの関心のある Web ページに加え、関心の無い Web ページも複数含まれていると考えられる。またこの方法では、閲覧履歴の取得、およびユーザへの新規ニュース記事の掲示のために専用のユーザインターフェースを使用しており、汎用性の面でも課題が残る。

ユーザの関心の取得に関連して、ユーザ間の共通話題によるつながりを発見するために、ユーザ各々のブックマークからユーザの関心を取得する研究[3]がある。この研究ではブックマークを「人手によってフィルタリングがかけられた、利用価値の高い情報である」と位置づけている。ブックマークに含まれる情報は、ユーザが過去に閲覧した Web ページ群の中から、ユーザ自身の手によって選別が行われたものである。

そこで本研究では、過去の閲覧履歴による選別ではユーザ関心に合致しないニュース記事を選別する可能性があったが、ブックマーク情報を利用することによりユーザの関心に合致したニュース記事をうまく選別できると考え、膨大なニュース記事の中からユーザ関心と類似度の高いものをユーザに掲示するフィルタの構築を行う。また、フィルタの入出力方法として RSS ファイルに着目し、汎用性の面でも工夫を加える。

2. システム概要

本研究では、特定の環境に限らず幅広い環境で利用できる汎用性を考慮し、新規ニュース一覧を RSS ファイルで取得し、フィルタ結果を RSS ファイルで出力するシステムを提案する。図 1 に本システムの概要を示す。

RSS（RDF Site Summary、RDF は Resource Description Language）は、Web サイトの更新状況を発信する仕組みである。RSS は Web ページのタイトルとリンクの一覧を記述したファイルであり、ニュース・サイトや日記サイトなど、更新頻度の高い Web サイトにおける更新状況の発信に利用されている。ユーザは RSS に対応した Web ブラウザ（RSS リーダ）を使用し、更新された Web ページのみ閲覧することができる。RSS リーダにはパソコンや携帯電話上で動作するものが複数存在する。

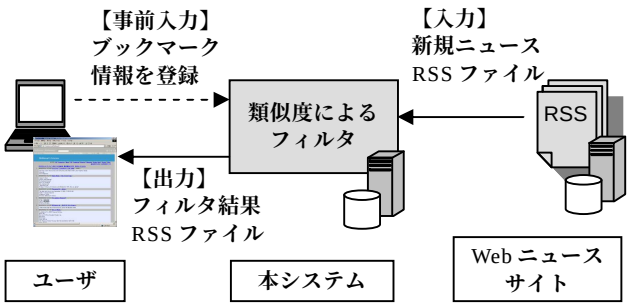

図 1 提案システムの概要

† 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 経営工学科

*1 http://www.asahi.com/
*2 http://news.ceek.jp/
*3 http://news.google.co.jp/

本システムでは、Web ニュース・サイトから RSS ファイルを取得し、リンク一覧の情報を元にニュース記事が記述された Web ページの html ファイルを取得する。さらに、ニュース・サイト毎に特有の html タグを基準としてニュース記事本文のみを抜き出して利用する。本システムをユーザ環境とは独立したサーバによるサービスとして実装・設置することにより、ユーザは、パソコンや携帯電話など任意の RSS リーダから利用が可能となる。

3. 提案フィルタ手法

3.1. フィルタリングの概要

本研究におけるフィルタリングの概要を図 2 に示す。処理内容は、事前処理・フィルタリング処理の 2 つに分けることができる。最初の Step では、事前処理として、ブックマークが指し示す Web ページの内容を元に、関心を表現するユーザプロファイルを作成する。この処理は本システムへのブックマーク情報の登録時に行われる。

次の Step では、RSS 情報のリンク先ニュース記事とユーザプロファイルの類似度を比較し、その値の大小により「関心あり」または「関心なし」のうち一方のカテゴリを付与する。全てのニュース記事を判定した後、「関心あり」カテゴリが付与されたニュース記事群を元に RSS を作成し出力することでフィルタリングが完了する。

以上の処理を提案法 1 とし、それに加えて提案法 2 では、事前処理において、過去のニュース記事から文書クラスタを作成し、プロファイル補充部がユーザプロファイルの補充処理を行う。

3.2. アルゴリズム（提案法 1）

本研究では、情報検索および情報フィルタリングの分野で広く用いられている、ベクトル空間モデルにおけるコサイン距離（文献[2][4][5]）を文書間の類似度の判断基準として使用する。文書 D に対して、D の単語を成分としてもつベクトルを単語ベクトル $\vec{t}=(t_i)$ とし、対応する単語の出現頻度 d_i を成分としてもつベクトルを文書ベクトル $\vec{d}=(d_i)$ と定義する。ここでブックマーク先 Web ページの文書ベクトル集合を $B=\{\vec{b}_1,\vec{b}_2,\cdots,\vec{b}_n\}$、フィルタリングを行う RSS 情報のリンク先ニュース記事の文書ベクトル集合を $R=\{\vec{r}_1,\vec{r}_2,\cdots,\vec{r}_l\}$、過去のニュース記事の文書ベクトル集合を $H=\{\vec{h}_1,\vec{h}_2,\cdots,\vec{h}_m\}$ とする。ただし単語ベクトルは集合 $R\cup B\cup H$ に含まれる文書の全ての単語を成分にもつベクトルとし、各文書ベクトルは出現しない単語に対応する成分を 0 とする。

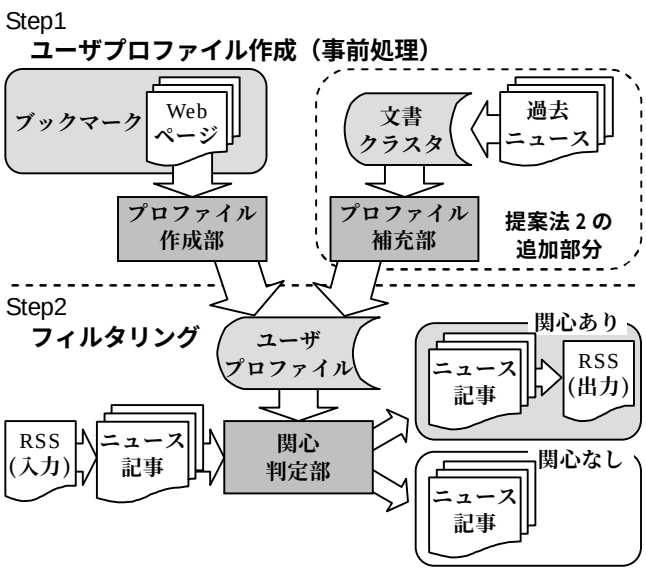

図 2　提案フィルタの概要

Step1：　（ユーザプロファイル作成）

① ブックマーク先 Web ページの内容を取得し形態素解析により単語分割を行い B を作成する。また、あらかじめ定期的に取得・蓄積した RSS 情報のリンク先ニュース記事から同様に H を作成する。

② 文書ベクトル集合に含まれるすべての文書ベクトル $\vec{d}_j=(d_j^i)$ に対して、下記の重み w_j^i を成分としてもつ重みベクトル $\vec{w}_j=(w_j^i)$ を計算し、重みベクトル集合 $W=\{\vec{w}_1,\vec{w}_2,\cdots,\vec{w}_{j\max}\}$ を作成する。集合 B に対応する重みベクトル集合を $P=\{\vec{p}_1,\vec{p}_2,\cdots,\vec{p}_n\}$ とし、これをユーザプロファイルと呼ぶ。

$$tf_j^i=\frac{\log_2(d_j^i+1)}{\log_2(tnum(\vec{d}_j))}$$

$$idf_i=\log_2\frac{(n+m)}{dfreq(t_i)}+1$$

$$w_j^i=tf_j^i\times idf_i$$

ここで、j は文書番号、i は単語番号、$tnum(\vec{d}_j)$ は文書ベクトル $\vec{d}_j$ 中の出現単語の種類数、n は B の文書数、m は H の文書数、$dfreq(t_i)$ は $B\cup H$ における単語 t_i を含む文書数である。また、tf_j^i は文書における単語の出現頻度を表し、idf_i は文書群における単語が出現する文書の偏りを表す。

Step2：　（フィルタリング）

① RSS 情報のリンク先ニュース記事を取得し、Step1 と同様の手順で R に対応する重みベクトル集合を作成する。なお、このとき R にのみ出現する単語の idf 値は 1.0 とする。2 つの文書 D_a と D_b の類似度は $sim(D_a,D_b)=\vec{w}_a\cdot\vec{w}_b/|\vec{w}_a||\vec{w}_b|$ により求める。類似度は 0 以上 1 以下の実数値をとり、値が大きいほど 2 つの文書は類似しているといえる。

② R の各文書に対して、B の文書との類似度が 1 つでも設定した閾値 s 以上の場合は「関心あり」のカテゴリを付与し、全ての類似度が閾値よりも小さい場合は「関心なし」のカテゴリを付与する。

なお、ユーザプロファイルはブックマーク先 Web ページそれぞれを 1 つのベクトルとした文書ベクトルの集合であり、その次元数は集合における総単語数である。おおよそ数千～数万次元の疎な行列となる。

3.3. 文書クラスタリングを利用したユーザプロファイルの補充

ブックマーク情報はユーザによって選別が行われたものである。その結果、ブックマーク情報にはユーザの関心の範囲に対してその一部の Web ページのみが含まれるが、実際のユーザの関心はもっと幅広いものである可能性が考えられる。例えば、ユーザが「自動車全般に関する話題」に関心があるが、一方でブックマークには「ホンダに関する話題」のみが登録されている場合を考える。

表１　フィルタ結果と正解カテゴリの比較方法

		提案法によるフィルタ結果	
		関心なしと推定	関心ありと推定(フィルタ出力)
正解カテゴリ	その他	f（正解）	α（誤選別）
	関心あり	β（選別漏れ）	t（正解）

このときフィルタ出力として得られるニュースは「ホンダに関する話題」であり、ユーザの関心のあるニュースであるものの、しかしユーザの関心の範囲である「自動車全般に関する話題」に対しては限られた範囲の話題でしかない。このような状況では、ブックマーク情報にある話題のみに限らず、ブックマーク情報に関連した幅広い話題をユーザは得たいのではないかと考えられる。

そこで本研究では文書クラスタリング[7]の利用に着目し、ユーザプロファイルに工夫を加えることを試みる。過去のニュース記事（集合 H ）に対して類似度によるクラスタ分析の手法を適用し、切断値 k を設定することで文書クラスタを作成する。作成された文書クラスタ群の中から、ユーザプロファイルとの類似度が高いものをユーザプロファイルに加えることで、自動的にユーザプロファイルの補充を行う方法を提案する。

3.4. アルゴリズム（提案法２）

3.2 節のユーザプロファイル作成後に、文書クラスタ作成処理とプロファイル補充処理を追加する。

Step1：　3.2 節の Step1 処理後

集合 H に対応する重みベクトルを作成し、類似度 sim 、切断値 k を用いてクラスタ分析を行う。次に B のそれぞれの文書毎に H との類似度を算出し、類似度が閾値 g 以上で最も高い文書 H_j の属するクラスタの全ての文書を B に追加して P を作成する。

Step2：　（フィルタリング）

提案法 1 と同じ処理を行う。

4. 評価実験

4.1. 実験概要

Web ニュース記事群（http://www.asahi.com/から取得）に対して提案手法を適用し、提案手法のフィルタリング性能を評価する。ニュース記事には、あらかじめ人手で関心ありニュース記事（関心記事）へのラベル付けを行い、正解カテゴリの集合を用意する。次に 2 つの提案法（3.2 節および 3.4 節）のフィルタ処理結果を正解カテゴリと比較する（表 1）。

得られるフィルタ結果と正解カテゴリがどの程度近いかを示す指標として、以下の評価基準を用いる。

$$\text{再現率} = \frac{t}{t+\beta} = \frac{\text{フィルタ出力のうち正解数}}{\text{正解カテゴリのうち関心記事数}}$$

$$\text{適合率} = \frac{t}{t+\alpha} = \frac{\text{フィルタ出力のうち正解数}}{\text{フィルタ出力総数}}$$

$$F\text{値} = \frac{2\times\text{適合率}\times\text{再現率}}{(\text{適合率}+\text{再現率})}$$

【前提条件】

① B ：関心のある任意の Web ページ 25 件

表２　提案法 2 における再現率・適合率

閾値 g	0.07	0.10	0.15	0.20
再現率	0.86	0.84	0.83	0.72
適合率	0.38	0.42	0.50	0.61

(a) s=0.07,k=1.00

切断値 k	0.07	0.10	0.15	0.20
再現率	0.84	0.84	0.84	0.84
適合率	0.35	0.35	0.40	0.42

(b) s=0.07, g=0.15

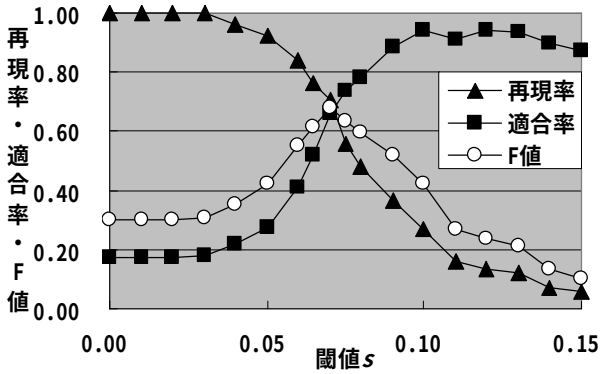

図３　提案法 1 における再現率・適合率・F値

② R ： Web ニュース記事 713 件
（期間：2005 年 10 月 20 日～10 月 26 日の 7 日間）
（関心記事 125 件、その他 588 件）

③ H ：Web ニュース記事 2690 件
（期間： 2005 年 9 月 22 日～10 月 19 日の 28 日間）

④ 単語ベクトルの構成要素：名詞、動詞、形容詞、副詞、未知語（文献[6]を参考に選択）

⑤ クラスタ分析方法：最長距離法(文献[7]を参考に選択)

⑥ 関心記事の判定方法：5 段階評価のうち「1 関心有り」のみを関心記事としてカウント

4.2. パラメータの決定

提案法 1 の結果を図 3 に示す。閾値 $s = 0.07$ 付近で再現率と適合率のグラフが交差し、このとき F 値は最大となる。以下では $s = 0.07$ を使用する。次に提案法 2 で使用する閾値 g および切断値 k の決定を行う。まず $s = 0.07, k = 1.00$ を設定し g を変化させたときの提案法 2 の結果（表 2(a)）から、再現率の変化量を考慮し $g = 0.15$ を選択する。次に $s = 0.07, g = 0.15$ を設定し k を変化させたときの提案法 2 の結果（表 2(b)）から、適合率の変化量を考慮して $k = 0.15$ を選択する。

4.3. 実験結果

提案法 1（ $s = 0.07$ ）、提案法 2（ $s = 0.07$, $k = 0.15$, $g = 0.15$ ）の比較を表 3 に示す。提案法１の適合率は 0.66 であり、ユーザの閲覧履歴を使用する先行研究[2]の適合率 0.64 と比較して同等の結果となった。これは、先行研究[2]では 5 段階評価のうち「1 関心あり」「2 やや関心あり」を共に関心記事としてカウントしているのに対し、本研究では「1 関心あり」のみを関心記事としてカウントする判定基準を使用しているため、適合率が低く算出されたためである。仮に先行研究[2]と同じ判定基準を使用すると、本研究における提案法１の適合率は 0.84 となる。ユーザ関心の取得に全ての閲覧履歴を使用する方法と比べ、ブックマーク情報を使用する方法が高い適合率が得られることを示唆する結果が得られた。

使用するブックマーク情報の件数 n とフィルタの選別性能の関係を図 4 に示す。n>15 で F 値が 0.7 程度となってお

表 3　提案法によるフィルタ結果

	提案法 1 (ﾌﾟﾛﾌｧｲﾙ補充なし)	提案法 2 (ﾌﾟﾛﾌｧｲﾙ補充あり)
再現率	0.70	0.84
適合率	0.66	0.40
ﾌｨﾙﾀ出力(t+α)	133	260
正解出力(t)	88	105
誤選別(α)	45	155
選別漏れ(β)	37	20

り、少ないブックマーク情報からでも高い選別性能が得られている。一方で、n>35 では n の増加に伴い再現率は向上するものの、適合率は低下している。これは、ブックマーク件数が多くなるほどフィルタ出力中の「1 関心あり」記事数が増加する一方で、「2 やや関心あり」記事数も共に増加し、それが「誤選別（α）」とカウントされ適合率が低く算出されてしまうためである。使用するブックマーク情報の件数を増やすほど、ユーザの関心に関連する記事が幅広く選別出力されている。

プロファイル補充の有無における比較として、n=25 に固定したときの、2 つの提案法における再現率と適合率の比較を図 5 に示す。適合率 0.40 以下、再現率 0.80 以上の範囲では、2 つの提案法はほぼ同じ選別性能であるが、それ以外の範囲では、提案法 1 の方が提案法 2 に比べ高い選別性能を示している。提案法 2 では、提案法 1 においてブックマーク件数を増加させた時と同様に、フィルタ出力中の「2 やや関心あり」記事数が増加しており、それが「誤選別」とカウントされ適合率が低く算出されている。これは、ブックマーク情報の件数が限られていても、提案法 2 のプロファイル補充を行うことによって、幅広くユーザの関心に関連する記事が選別出力される効果が期待できることを示している。

表 3 におけるフィルタ結果では、提案法 1 に比べ提案法 2 は、選別漏れは半減している。その減少件数 17 件のうち 4 件の記事は、提案法 1 で閾値を s=0.05（再現率=0.92、適合率=0.28）まで低くしても出力されない記事であった。提案法 2 では、適合率=0.40 において、これらの記事を選別することができている。提案法 2 では、提案法 1 では選別しにくい記事を選別することができており、2 つの提案法では異なった性質を持つことが示されている。一方、誤選別は 3 倍強に増加しているが、その中には「やや関心あり」記事のほか、「関心なし」記事も選別出力されており、それを減らすことは今後の課題である。

5. おわりに

本研究では、ブックマーク情報に基づいたニュース記事のフィルタとして、２つの手法を提案した。実験により、提案法 1 では先行研究と同等以上の適合率が得られ、ユーザの関心に合わせたニュース記事の選別が行われる事を示した。提案法 2 では、さらに再現率が向上しユーザの関心のあるニュース記事が正しく選別される一方、適合率が低下することが分かった。

今後の課題として、プロファイル補充ありの場合における「関心なし」記事の誤選別を防ぐ必要がある。その解決手段の１つとしては、単語の共起情報の利用（文献[6][7]）が考えられる。

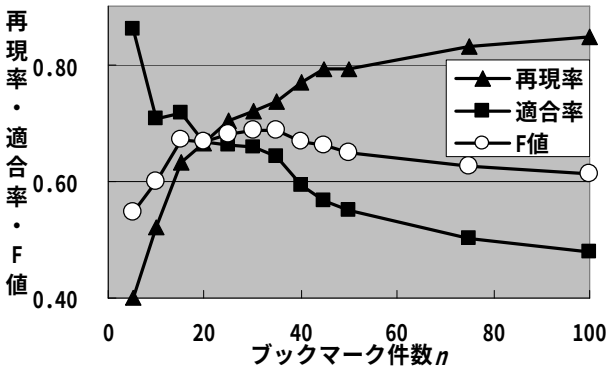

図 4　ブックマーク件数と選別性能

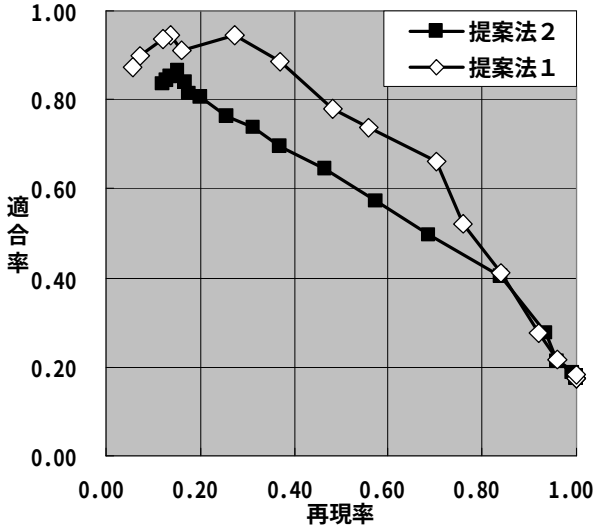

図 5　2 つの提案法における選別性能の比較

参考文献

[1] 白井康之，情報フィルタリングの双対性，三菱総合研究所 週刊 Take IT Easy http://easy.mri.co.jp/20050405.html

[2] 向井誠,青野雅樹（2005/09），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，情報処理学会研究報告 Vol.2005,No.94,pp.27-pp.32

[3] 濱崎雅弘,武田英明,松塚健,谷口雄一郎,河野 恭之,木戸出正継（2002），Bookmark からの共通話題ネットワークの発見手法の提案とその評価，人工知能学会論文誌 Vol.17,No.3,pp.276-284

[4] 金明哲,村上征勝,永田昌明,大津起夫,山西健司（2003/03），言語と心理の統計，pp.87-92 岩波書店

[5] 中川裕志，ベクトル空間モデル，東京大学 大学院情報学府 情報データベース論 講義資料, http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/

[6] 松尾豊,石塚満（2002/05），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，人工知能学会論文誌，Vol.17,No.3,pp.217-223

[7] 小熊淳一,内海彰（2005/06），語の共起情報を用いた文書クラスタリング，人工知能学会全国大会第 19 回, http://www-kasm.nii.ac.jp/jsai2005/schedule/paper-91.html