

中平 頼孝[†] 千種 康民[†]

[†]東京工科大学大学院 バイオ・情報メディア研究科

1 はじめに

情報が溢れている現代、多数の Web サイトの更新情報を得るには時間と手間がかかる。これを解決するために登場した技術が RSS である。しかし、近い将来ユビキタス社会の到来により、さらに多くの情報が溢れ、多くの RSS を管理していく時代が予想される。このような時代になったとき RSS を管理し、目的別に整理する作業は非常に煩雑である。本研究では、パーソナリティにあった RSS を配信することにより将来発生するであろう、この問題を解決する。

2 提案手法

本研究では、RSS の特徴語抽出とユーザ履歴から興味語抽出を行う。この結果を元に、ユーザに提示すべき RSS かの推定を行う。ユーザに提示した RSS の履歴、閲覧履歴などは全て、次の提示を行う際の情報としてユーザ履歴に保存する。本システムの流れを図 1 に示す。

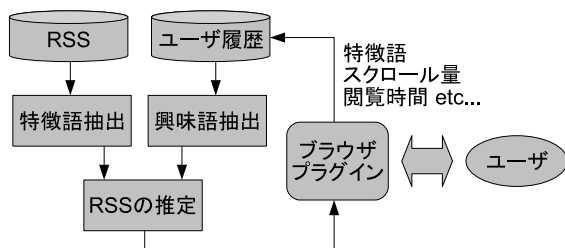


図 1: システムの流れ

2.1 RSS からの特徴語の抽出

出単語の重要度を評価する TF・IDF 法を用いて、RSS の特徴語を抽出する。

TF_i 対象となる文書内における総単語数 (T) と、その単語 (i) が出現する頻度 (r_i) の割合を示したもの

$$TF_i = \frac{r_i}{T} \quad (1)$$

IDF_i 比較文書数 (N) の中で、その単語 (i) が出現する文書数 (n_i) が少ないほど、その単語が文章を特徴付けている

$$IDF_i = \log \frac{N}{n_i} \quad (2)$$

Delivery of RSS in accord with personality

Yoritaka NAKAHIRA[†], Yasutami CHIGUSA[†]

[†]Tokyo University of Technology, E-Mail chigusa@cc.teu.ac.jp

URL <http://www.teu.ac.jp/chiit/>

2.2 クリックによる興味度の算出

商品などの物に関しては、購入履歴から、その物に対して、興味を持っているという強い信号を受け取れる。しかし RSS は情報であり、本当にその情報を取得しているのかどうか、本当に情報を読んでいるのかどうか、と言うように、興味を持っている情報に対しての信号を受け取ることが難しい。

そこで本研究では、2 つの仮説を立て、ユーザが興味を持つ単語を抽出する手法を提案する。

仮説 1 いま読んでいる RSS^q の記事内にあるリンクをクリックした時、クリック文章内に RSS^q の Title に含まれている単語があった場合、その単語にに対して興味を持っている。

この仮説に従い、 RSS^q についてタイトルとクリックした文章に含まれる単語 (i) に対する興味度 (wp_i) を算出する。

$$wp_i = \frac{r_i^{RSS}}{T^{RSS}} \times \frac{r_i^{click}}{T^{click}} \times TF_i^q \cdot IDF_i^q \quad (3)$$

r_i^{RSS} : RSS^q のタイトル内での出現頻度

r_i^{click} : クリック文章内での出現頻度

T^{RSS} : RSS^q の総単語数, T^{click} : クリック文章の総単語数

RSS^q : ユーザが閲覧した RSS

仮説 2 RSS 記事内の「次」や数字を含む文章 (2 ページ目など) のリンクをクリックした時、 RSS^q に興味を持ってユーザが閲覧したと言う強い信号である。

2.3 閲覧時間とスクロールに応じた興味度の算出

仮説 2 に当てはまった時、ユーザの閲覧時間 (time) とページのスクロール (scroll) の関係は図 2 のような傾向を取る事が分かった。

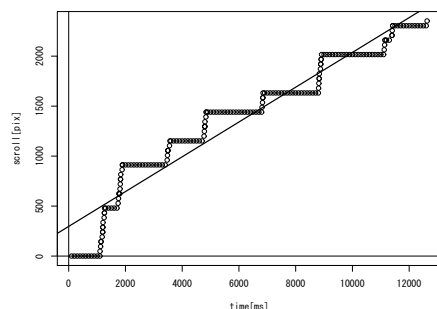


図 2: ユーザの閲覧情報

この傾向を利用し、仮説 2 に当てはまるデータ群に対して一次式を最小二乗法により求める．この一次式の傾きをユーザの閲覧速度 (u) として、 RSS^q を興味を持って閲覧したかどうかの評価に使用する．

仮説 2 に当てはまらない RSS^q の閲覧速度を v としたとき、単語 (i) に対する興味度 (op_i) を求めることができる．

$$op_i = (TF_i^q \cdot IDF_i^q + \alpha wp_i) \times \frac{u}{v} \times K^q \quad (4)$$

K^q : 閲覧の割合によって決まる係数
 α : スケーリング係数

2.3 推薦する RSS の推定

RSS がユーザにとって適した情報かどうかの判定を行うために、 RSS^p の特徴語を用いて、 RSS^p に対するユーザの興味度を推定する． RSS^p 内に現れる全ての特徴語と、一致する興味語に対して補正を行い、ユーザに推薦すべき RSS の推奨度を求める．

$$occ(RSS^p) = \sum_{i=w}^M (TF_i^p \cdot IDF_i^p \times op_i)$$

ただし、 $w \in G(RSS^p)$

RSS^p : 判定の対象となる RSS
 $occ(RSS^p)$: RSS^p の推奨度
 M : RSS^p 内に現れる特徴語の個数
 $G(RSS^p)$: RSS^p 内に現れる特徴語の集合

3 評価

3.1 実験内容

ユーザに本システムを使用して、提示される RSS の推奨度 (occ) とユーザの単語に対する興味度 (op_i) に対して評価を行う．毎回、推奨される RSS は 30 件、興味語は上位 100 件で評価を行い、計 10 回試行する．

なお、実際のシステムでは日付のポイントを入れ、推奨順に表示するが、実験ではシステムがどのような推定を行っているのかを知るため、推奨順でのみ RSS を抽出する．

3.2 実験結果

表 1: 実験結果

回数	推奨度 [%]	興味語の正当性 [%]
1	10.0	-
2	46.7	6.0
3	23.3	23.0
4	20.0	30.0
5	40.0	35.0
6	50.0	44.0
7	43.3	51.0
8	40.0	52.0
9	30.0	51.0
10	43.3	57.0

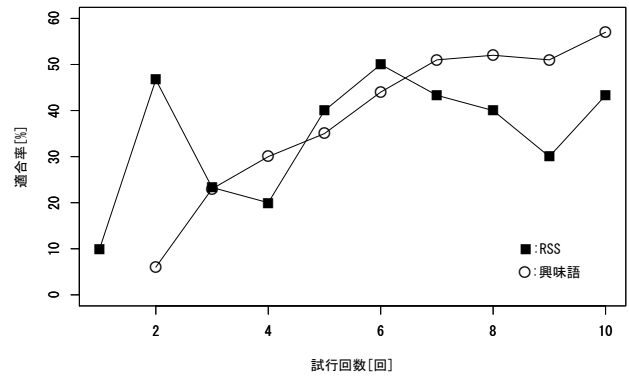


図 3: 試行回数と適合率

3.3 考察

現状では、ある程度学習を行った後の推奨度は 30 ~ 50% であり、興味のある単語の正答率は 50% である．実験では、興味が無くても推奨されてしまう RSS が、永続的に抽出されてしまうという問題点があった．実験の後半の RSS 適合率が上昇しなかったのは、これらの RSS が目立った為だと考えられる．

4 おわりに

図 2 は、一般的なユーザの閲覧情報として例を挙げ、仮説を立てた．しかし、ユーザが閲覧するときの癖などが大きく影響を及ぼす可能性がある．これらを検証する為、多くのユーザに使用してもらい、多様な傾向を導き出す必要がある．導き出した情報を元に、ある傾向が多く見られたとき、システムが自動的に学習し、閲覧したかどうかの評価を行う機能を実装する予定である．

実験で問題となった永続的に抽出されてしまう RSS に対しては、提示してるにも関わらず、いつまでも閲覧されない RSS に対して、マイナスの興味語となる「無興味語」を抽出することにより、問題は解決と思われる．

今後は、このルールの適応と、興味語と無興味語とのバランス調整が課題となる．

参考文献

- [1] 中平頼孝, 松永雄平, 福元雅司, 千種康民, 大野澄雄, 小池隆: TPO とユーザプロファイルに基づいた情報提供システム開発, 情報処理学会第 66 回全国大会論文集デ-05
- [2] 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, 情報処理学会研究報告 2005-FI-80(5)
- [3] 堀江育也, 加瀬誠志, 北守一隆: エモーショナル・センシングを用いた Web 診断, 経営情報学会 2003 年度