

RSS を用いたサイト登録支援システム

滝沢 敏裕[†] 湯浅 将英[‡] 大山 実^{†‡}

東京電機大学 情報環境学研究科 情報環境工学専攻[†]

東京電機大学 情報環境学部 情報環境工学科[‡]

1. はじめに

近年 RSS が注目されており、多数の Blog やニュースサイトでは RSS を配信している。RSS を有効に利用するためには専用のリーダ（RSS リーダ）を用いる。しかし、多くの RSS リーダでは閲覧サイトが配信している RSS を探し出すための手間がかかる。そこでランキング表示や検索、推薦機能によって RSS の登録を支援システムの作成した[1][2]。本稿ではさらに、RSS の文書を解析することで得られる情報をもとにしたカテゴリの自動分類と、その予備実験の結果を述べる。

2. 従来まで開発したシステムの概要

本研究で開発したシステムはサーバサイド型で、ユーザの RSS 利用情報を蓄積し、各ユーザに RSS の情報をランキング、検索、推薦の形で提供する。これらによって、ボタンひとつで RSS を登録することができる。

2.1 システムの機能

① RSS 利用状況の提供

システムを利用しているユーザが登録している RSS を集計することでランキングを作成する。これにより、人気のある RSS やユーザが興味のある RSS を簡単にみつけることが出来る。

② 記事の検索

ユーザが登録している RSS の記事はシステムによって収集され、検索が行えるようになる。これによりユーザはキーワードをもとに、特定の情報に関する記事を配信する RSS をみつける。

③ カテゴリを用いた推薦

嗜好情報を利用した推薦機能について説明する。システムにはそれぞれユーザが登録している RSS の情報が保存されている。登録された RSS にたいしてユーザはカテゴリを定義できる。これらの登録された情報をベクトル情報に変換し、協調フィルタリングにより RSS を推薦する。

2. 2 問題点

カテゴリを用いた推薦ではユーザ自身が RSS にカテゴリを定義する必要があり、これはユーザの負担となる。そのためユーザによるカテゴリ定義が行われないことがあり、システム全体の推薦の精度が下がってしまう可能性がある。筆者らが開発したシステムも、カテゴリが定義されなければ推薦がうまく機能をしない場合がある。そこでユーザによるカテゴリ定義だけでなく、カテゴリ定義の自動化をする機能があることが望ましい。

3. RSS を用いた文書解析

3.1 RSS の情報を利用

本研究では RSS に対して自動カテゴリを定義することを目指し、RSS の文書を解析することで特徴を抽出する。そして RSS ごとの類似を比較し、分類をする。

3.2 RSS 文書の解析

Web の文書から文書解析によって特徴を得るためには、ノイズである情報を除外し本文のみを解析する必要がある。そのため一般の Web の解析は時間がかかってしまう。一方、RSS は構造化された文書であり、本文抽出などの処理を必要としない。ただし RSS は要約された情報を扱うことを主旨として配信されているため、情報量の多いものや少ないものが混在している。最も情報量の多い RSS では本文の内容全てを含んでいる。また、最も少ないものはタイトルとリンクのみとなっている。ここでは RSS のタイトルを用いたカテゴリ分類を考える。しかし RSS のタイトルでは、従来の本文抽出を用いた文書解析よりも利用できるデータが少なくなってしまう、自動カテゴリの定義の精度が下がってしまう可能性もある。RSS のタイトルみを用いてカテゴリ分類が正しく行えるか否かの実験をする

4. 予備実験

実験ではあらかじめ筆者が内容を事前に吟味し、国際、スポーツ、芸能、社会、経済、政治、ゲームの 8 つのカテゴリに定義した約 40 個の RSS を用意した。また、文書解析の対象データは条件を合わせるため RSS が配信する記事のタイトルのみを使用した。実験において文書解析（形態素解析）には chasen を使用した。また、重み付けのための TF-IDF の算出には有名検索エンジンにより提供される API を用いその値をもとに IDF を求めた。

Bookmarking Assistance System using RSS

Toshihiro TAKIZAWA[†],

Masahide YUASA[‡] and Minoru OHYAMA[‡]

[†]Graduate School of Information Environment Tokyo Denki University.

4.1 自動カテゴリ分類

それぞれの RSS が配信する記事一週間分のタイトルに形態素解析を用いて、品詞に数を含まない名詞や未知語などの単語を抽出し、それぞれの単語の TF-IDF を求める。この TF-IDF の値をもとにコサイン相関値による RSS 同士の類似度を算出する。これにより類似度が、想定したカテゴリの情報に基づいているか否かを調べる。

4.2 実験手法

実験に用いたそれぞれの RSS に対して類似度を算出する。類似度の平均値、平均値の 2 倍、2.5 倍、3 倍を基準値とし、基準値を上回る値を持った RSS を類同一カテゴリとして採用する。これらの RSS のカテゴリが正しいか否か、適合率と再現率を求める。なおカテゴリが正しいか否かは人手によりサイトの内容を確認して判断した。

4.3 実験結果

実験ではデータの欠損があったもの、カテゴリの定義が 8 種に該当しないものを除く 36 個の RSS を最終的に使用した。表 1 は基準値ごとの適合率と再現率を表したものである。

表 1 基準値ごとの適合率・再現率

基準値	適合率(%)	再現率(%)
平均値	33.81%	92.56%
平均値*2	53.09%	79.69%
平均値*2.5	68.96%	62.50%
平均値*3	62.5%	54.69%

表 2. RSS「国際のトピックス-goo」の類似度

RSS 名	類似度
国際のトピックス - goo	1.00
YOMIURI ON-LINE / 国際	0.57
Sankei Web / 国際	0.48
読売ニュース / 国際	0.42
livedoor News 海外	0.38

表 3. RSS「livedoor News 経済の類似度」の類似度

RSS 名	類似度
livedoor News 経済	1.00
CNET Japan	0.24
Itpro 国内ニュース	0.24
読売ニュース / 経済	0.23
MYCOM ジャーナル	0.21

5. 考察

表 1 は基準値ごとの適合率と再現率を示したものがある。基準値を何にするかで適合率や再現率の値が変わるため、今後、どのような基準値を設定すべきかが課題となる。表 2 は最も良く分類の出来ている RSS の類似度の表である。RSS「国際のトピックス-goo」は国際に関連する内容をもった RSS であり、そのほかの RSS もすべて国際に関連する内容を持ったものである。表 3 はもっとも分類の出来なかった RSS の類似度の表である。RSS「livedoor News 経済」は経済に関連する内容を持つ RSS であるが、CNET Japan や Itpro といったコンピュータに関連する内容をもった RSS が類似している。表 3 のような結果になった理由としては、livedoor News 経済はコンピュータ関係の内容を含んだ経済のニュースを配信していた。このことにより、コンピュータに関連する RSS が類似していると分類された。これらの結果から、従来のように Web の本文抽出による解析結果から特徴的な単語を探すのではなく、RSS のタイトル情報にある程度集め、文書解析し特徴的な単語を探し出すことでも、自動分類がほぼ出来る見通しが立った。

6. まとめ

RSS の登録を支援するシステムにおけるカテゴリ定義を目標に、RSS のカテゴリ分類を試みた。予備実験より RSS のタイトル情報を用いることで RSS のカテゴリ分類が出来る見通しが立った。今後は RSS の数を増やした本格的な実験をし、カテゴリの自動分類を用いたカテゴリの自動定義を目指す。

参考文献

- [1] 滝沢敏裕, 山名健悟, 湯浅将英, 大山実: “複数ユーザの利用情報を用いた RSS リーダ”, 情処学, 第 67 回全国大会, 3, pp.197-198(2005)
- [2] 滝沢敏裕, 山名健悟, 湯浅将英, 大山実: 閲覧サイトの登録を支援する RSS リーダ, FIT2005
- [3] 山名健悟, 滝沢敏裕, 湯浅将英, 大山実: blog 検索による商品選定支援システム, 情報処理学会第 55 回 デジタルドキュメント研究会, (2006)