E-011

記事群の冗長度を削減するためのRSSニュースリーダ

Attaching Reducing Articles Redundancy Facility to RSS News Reader

中渡瀬 秀一† 戸田 浩之† 片岡 良治†
Hidekazu Nakawatase Hiroyuki Toda Ryoji Kataoka

1．まえがき

近年，インターネットにおけるニュース配信の形態として，RSS形式のデータによる提供が盛んになっている．このRSSにはWebサイトの更新情報として記事タイトル，その URL，更新時刻などが含まれている．そしてこのRSS データを閲覧するためのユーザアプリケーションがRSS(ニュース)リーダである．ユーザは RSS リーダに関心のあるサイトの RSS アドレスを登録しておくと，RSS リーダは定期的に RSS の更新を監視し，サイトの新着記事をタイムリーに知ることができる．また RSS リーダに複数のサイトの RSS を登録しておくことにより，ひとつのアプリケーション上で同時に複数サイトの更新や新着情報を容易に確認することができるのがその利点である．しかし複数ニュースサイトの記事をひとつの GUI 上に集めると，異なる報道元から同じ事件に関するほぼ同一内容の記事が多数重複していることが分かる．事件の事実内容だけを確認したい場合にはこれらのうちどれかを読めば十分であるため冗長さが問題となっていた．

我々はこの問題を改善するために同一事件に関する記事を集約することができるような RSS リーダシステムを試作した．

2．RSSニュースリーダシステムの概要

本システムは Web サーバ部と Web クライアントからなる Web 上のアプリケーションである．

2．1 RSSニュースリーダシステムの構成

システムの全体構成を図 1 に示す．システムは以下の部分から構成される．

- RSS クローラ
- 記事データベース
- 記事集約制御部
- GUI 生成部
- Web サーバ
- Web クライアント（ブラウザ）

RSS クローラは登録された RSS の URL から定期的にニュースサイトの RSS を取得して新着ニュース記事の有無(サイト更新)を監視する．もし新着記事が存在すればその情報(タイトル，配信時刻，記事 URL)を記事データベースに格納する．

記事集約制御部では記事データベースに格納された記事を比較して同一事件に関する報道記事をグルーピングする．グルーピングの詳細については2.2節で説明する．

GUI 生成部は複数のニュースサイト RSS から収集した記事タイトルを集約されたグループ単位で閲覧できるような GUI を生成する．GUI は図 2 に示すように 2 ペインから構成され，右側に集約された記事グループの一覧が代表タイトル(グループ中の任意の記事タイトル)の時系列で表示され，タイトルをクリックすると元記事が左ペインに表示される．またフォルダアイコンをクリックするとグループ内に含まれる記事タイトル一覧と元記事へのリンクが表示される．

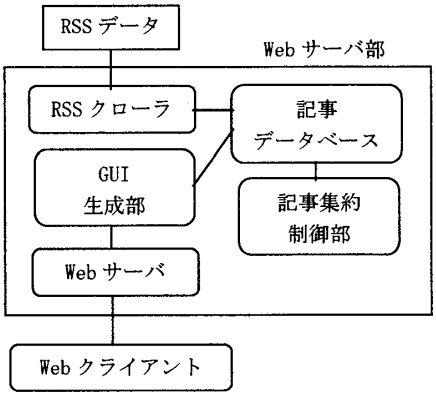

図 1：RSS ニュースリーダシステム

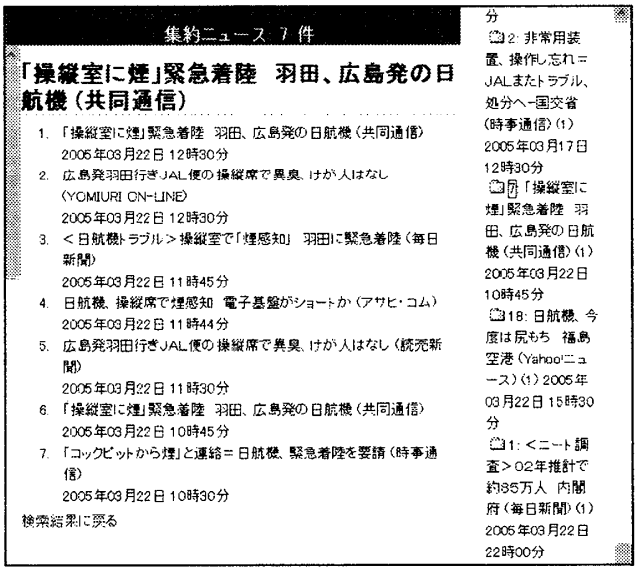

図 2：Web クライアント概観

2．2 同一事件記事集約方法

与えられた 2 記事が同一事件に関する報道記事であるかどうかを判断するために本システムが使用可能な情報はRSSから得られる記事タイトルと配信日時である．したがってタイトルの類似度が高く，双方の配信日が同じ 2 本の記事は同一事件の記事とみなすこととした．

タイトルの類似度を測るのには従来の代表的手法であるベクトル空間[2][3]に基づいた手法を用いた．これは文から特徴語を抽出して，それらの組による特徴語ベクトル

†日本電信電話株式会社 NTTサイバーソリューション研究所

の演算によって類似度を計算するものである．処理手順の概略を以下に説明する．

(1) 特徴語の抽出

特徴語を抽出するためにまず記事のタイトル文を形態素解析する．解析器には日本語形態素解析システムJTAG(α-Tagger)[1]を用いた．これによってタイトル文を分かち書きして形態素に分解する．その結果得られた形態素の品詞属性が名詞もしくは未知語(ただし記号類は除く)であるものを特徴語として抽出する．

(2) 特徴語の正規化

表記が異なるが意味の等しい特徴語を表現の正規化によって統一する．ここでは漢数字とアラビア数字の表現を統一するなどの処理を行う．

(3) 特徴語ベクトルの類似度計算

抽出された特徴語の組によるベクトルがその記事を表現する特徴語ベクトルとなる．このときベクトル間の類似度をそれらベクトルが共有する特徴語の数(ベクトルの内積に相当)とする．この値が大きいほど類似度が高い．以下に計算例を示す．

(A)"ヤフーが JASRAC と契約締結、音楽無料配信サービスを拡大へ"

(B)"ヤフーと JASRAC が音楽ネット配信契約、610 万曲利用可能に"

文(A)，(B)はある記事のタイトル文である．これを JTAG で形態素解析した結果を特徴語ベクトルに変換すると以下のようになる．

(A') (ヤフー，JASRAC，契約，締結，音楽，無料，配信，サービス，拡大)

(B') (ヤフー，JASRAC，音楽，ネット，配信，契約，610，万，曲，利用)

この場合，(A')と(B')が共有する特徴語は

{ヤフー，JASRAC，契約，音楽，配信}の 5 個であるのでこの場合の(A)と(B)の類似度は 5 となる．

(4) グルーピング

2 本の記事を比較する際，類似度に閾値を設けて，それらの類似度がその閾値以上であるならば 1 つにグルーピングする．そのためにあらかじめ任意の 2 本の記事について類似度を計算しておく．これによってある記事と集約されるべき他の記事を決定することができる．この閾値が小さいほど集約の効果が高い．しかし小さすぎると同一事件でない記事がグルーピングされることがあり，また大きすぎるとグルーピングされた記事であれば同一事件に関するものである可能性は高くなるが一方，同一事件にもかかわらずグルーピングできないものも増加する．

3. 使用結果

朝日新聞，日本経済新聞など大手全国紙やスポーツ誌，コンピュータ専門誌が運営するサイトなどから 20 サイトの RSS を選び本システムに登録してシステムの運用実験を行った．期間は 2005 年 3 月 1 日～3 月 20 日の 20 日間である．その結果を表 1 に示す．実験の結果，1 日当たり約 2200 記事を収集することができた．これを集約して同一事件記事のグループに集約すると平均約 1250 グループであった．この集約で 1 グループを記事 1 本に換算することによって約 42%の記事数を削減したことになる．また集約グループのサイズとその度数との関係も調査した．その結果，2 本以上の記事が集約されてグループ化されたケースが 1 日平均 300 グループ以上，5 本以上でも 60 グループ以上あることが確認された．この分布状況を図 3 に示す．今回の集約結果を見ると図 2 に見られるように同一事件の報道元による表現の差異にかかわらず集約が成功している例は多かった．しかし過度に集約されてグループのサイズが 70 を超えるような例も見られた．以下はそのような記事グループに含まれる 3 記事のタイトルである．

"松井稼は 3 打数 1 安打　大リーグオープン戦"

"田口、1 安打と盗塁　大リーグオープン戦"

"松井秀、右翼に 3 点本塁打　大リーグオープン戦"

この 3 記事の内容は厳密には別の事実である．同様の例としては 1 日に類似した短いタイトルの記事が何本も配信される市況ニュースが挙げられる．

	値
新着記事数(総数) A	43474
新着記事数(1 日平均)	2174
集約グループ数(総数) B	25020
集約グループ数(1 日平均)	1251
B/A	0.575516

表 1：記事数など(2005 年 3 月 1 日～20 日)

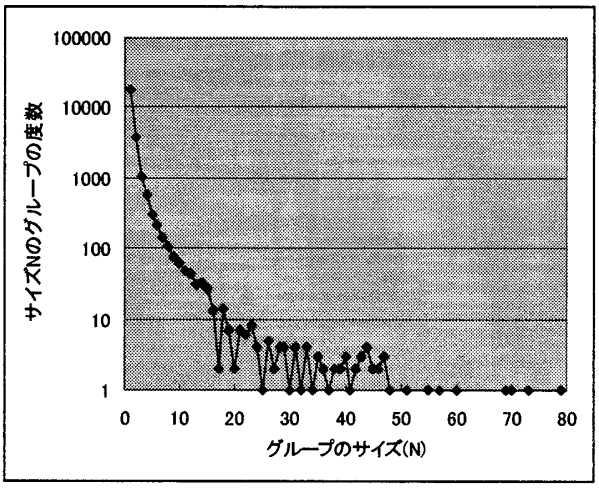

図 3：集約グループサイズとその度数

4. まとめ

我々はニュース記事の冗長度を記事のグループ化によって削減できる RSS ニュースリーダの設計と構築を行い，その動作を確認した．その結果，RSS 情報から新着記事を収集し同一事件記事を集約することができた．今後の課題として集約精度の評価，誤集約の改善などが挙げられる．また記事グループのサイズは事件の注目度に相当することからこの利用も検討したい．

参考文献

[1] 日本語形態素解析システムα-Tagger, http://www.nihongo-solution.info/product/alpha_tagger.html

[2] 徳永健伸: "情報検索と言語処理", 東京大学出版会 (1999)

[3] W. B. Frakes and R. Baeza-Yates Eds.: "Information Retrieval: Data Structures & Algorithms", Prentice Hall(1992).