Vol.2009-IS-110 No.6
2009/12/1

インターネットコミュニティ活性化を目的とした ウェブ記事収集システムの提案

鈴木康祐† 岡本東† 堀川三好† 菅原光政†

本研究では，企業・自治体・地域事業といったコミュニティ向け CMS の支援を目的としたニュースサイトを構築する．提案システムでは，CMS を基に作成されたプロファイルに基づき，ニュースサイト運営者に候補記事を提示する．また，記事収集の精度向上のため，適合性フィードバックを用いたプロファイルの更新手法を提案する．これにより，コンテンツの充実化による新規利用者の獲得，サイト活性化効果による継続的な運営を実現する．

Development of Article Collection System to activate for Internet Community

Kousuke Suzuki†, Azuma Okamoto†, Mitsuyoshi Horikawa† and Mitsumasa Sugawara†

In this paper, we constructed the News Site in which it aimed at the support of CMS for the communities such as an enterprises, municipalities and regional business. The proposed system presents some similar articles related to the profile based on CMS. In addition, it proposes the technique to improve the Article Collection System by using the relevance feedback. As a result, continuous management by enhancing some contents in order to get new users and effect of the site activation is achieved.

1. はじめに

近年，SNS (Social Networking Service) やブログといった，インターネット上での情報発信・共有を行うことを目的とした CMS (Content Management System) が注目を集めている．それに伴い，CMS を基盤としてインターネット上で交流を行うインターネットコミュニティも増加している．

本研究では，企業・自治体・地域事業といった実社会の組織を基にしたコミュニティを対象とし，CMS における情報収集・コミュニケーションの活性化を目的としたニュースサイトを構築する．これにより，コンテンツの充実化による利用者の拡大，サイト活性化効果による継続的な運営の実現を目的する．

提案するニュースサイトでは，時事性の高いニュースだけではなく，ウェブ検索によって得られた専門性の高いブログやホームページを収集する．しかし，記事の膨大さやコミュニティの興味の不明確さから，運営者だけで適切なニュースを継続的に提供することは難しい．そこで，インターネット上のコミュニティが交流の基盤とする CMS から，利用者の興味を示すプロファイルを作成し，それらに基づき外部サイトからウェブ記事を収集して提供する手法を提案する．その際，作成されたプロファイルと候補ウェブ記事をベクトル空間モデルに基づき比較を行い，算出された類似度を基にランキングすることで，運営者が提供するウェブ記事の選択の支援を実現する．

また，提案した手法を，学童保育事業向け CMS である岩手県学童保育情報サイトに適用した数値実験を行い，有効性を検証する．さらに，実験結果から判明した問題点を改善するため，適合性フィードバックを用いたプロファイルの更新手法を提案する．この手法では，ランキング上位のウェブ記事に対して運営者が分類した適合・不適合文書を基にプロファイルを更新し，候補となる記事を再提示することで，ニュースサイト運営者の負担の軽減を実現する．

2. ニュースサイトによるインターネットコミュニティの活性化

2.1 対象とするインターネットコミュニティ

インターネット上で交流を行うインターネットコミュニティは，様々な観点から分類・整理される．小笠原 [1)]はその一例として，①母体が現実社会の既存集団か，②インターネットで初めて形成された集団か，といった分類方法を提案した．本研究では，①の実社会を基にしたコミュニティの中でも，企業・自治体・地域事業といった大規模なコミュニティを対象として想定する．企業・自治体・地域事業といったコミュニティの大きな特徴として，組織の活動支援を目的としているため，より継続的な運営

† 岩手県立大学大学院ソフトウェア情報学研究科
Iwate Prefectural University Graduate School

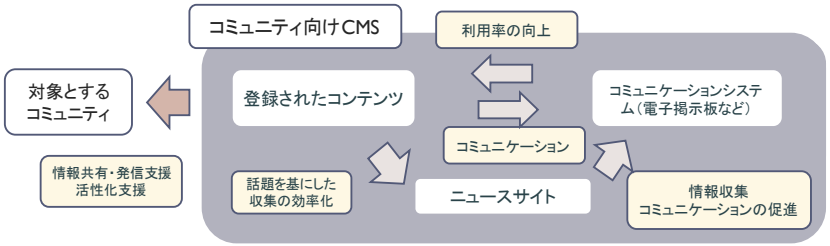

図 1　コミュニティ向けニュースサイトの概念図

が求められる．そのため，運営者が主体となり，情報収集の効率化・コミュニケーションの促進を実現していく必要がある．

2.2 ニュースサイトとは

ニュースサイトとはインターネット上にてニュースを掲載・提供しているウェブサイトの総称であり，新聞社などが自社のニュースを提供するサイトと，そのようなメディアと提携してニュースを提供するサイトに大別される．また，個人が運営するパーソナルニュースサイト[2]が存在する．特徴として，メディアのニュース記事を引用するだけではなく，特定の分野に特化した専門的なホームページやブログを紹介するケースが多く見受けられる．

本研究で提案するシステムに類似した形態をもつサイトの事例として，国内では newsing[3]や Buzzurl[4]，海外では digg[5]など，情報収集・コミュニケーションを促進するための仕組みを有したサイトが挙げられる．このようなサイトでは，利用者によるコメントや評価・議論の仕組みといった，利用者同士の交流を図るための機能が充実化されており，ニュース記事を通した意見交換・議論から，利用者間のコミュニケーションを促進する効果が期待できる．

本研究では，インターネット上で形成されるコミュニティではなく，実社会を基にしたコミュニティを対象として想定している点に特徴がある．前述したパーソナルニュースサイトのように，特定の分野に特化した記事を主体に収集・提供することで，コミュニティにおける情報収集・コミュニケーションを支援できると考えられる．

2.3 インターネットコミュニティへの効果

本研究では，コミュニティ向け CMS における継続的な運営を図る手段として， ニュースサイトにおける高い情報収集能力と，それらを基にした議論・意見交換などコミュニケーションを促す仕組みに着目する．コミュニティが交流を行うための基盤であるコミュニティ向け CMS に対してニュースサイトを導入することにより，以下の効果が期待できる（図 1）．

(1) 情報の充実化による利用者の獲得

コミュニティに関連するウェブ上のニュースを中心に収集・提供することで情報の充実化を図り，新規利用者や主体となる利用者を獲得できる．

(2) コミュニケーションの促進による活性化

提供されたニュースに関する話題について議論する仕組みを作ることで，利用者同士のコミュニケーションの活性化を図る．

3. コミュニティ向けCMSに基づくウェブ記事収集システム

3.1 問題点

提案するニュースサイトでは，大手の新聞社が配信するニュースだけではなく，専門的なホームページやブログといった幅広い記事を提供する．時事性の高いニュースは，RSS (Rich Site Summary) といったツールの利用や他のニュースサイトの閲覧によって入手できるが，コミュニティに特化した専門性の高い記事をウェブ上から収集する際には，以下の問題点が挙げられる．

(1) ウェブサイトの膨大さ

ホームページやブログなどのウェブ記事はウェブ検索で探す必要があり，運営者の負担となる．

(2) ニュースとして適合性が高い記事の判別

ニュースサイトでは，提供するべきウェブ記事の選択は運営者の主観に依るものが大きい.そのため，ニュースとしての適性が高い記事の見落としや，ニュースとしての適性が低い記事の誤選択が生じる．

3.2 提案するシステムの概要

3.2.1 概要

上記の問題点を解決するため，特定のコミュニティのためのウェブ記事収集を支援する仕組みを提案する．提案するシステムでは，コミュニティがインターネット上で利用する CMS 内にて話題となっている内容に利用者は興味を持っていると仮定し，外部ウェブサイトからウェブ記事を収集して提供する．その際に，CMS から利用者の嗜好を示すプロファイルを定義し，候補記事との類似度をベクトル空間モデルに基づき算出・ランキング形式で出力することで，提供する記事の選択支援を図る．

3.2.2 対象とするCMS

CMS とは，ウェブコンテンツを構成するテキストや画像，レイアウト情報などを一元的に保存・管理し，サイトを構築・編集するソフトウェアのことで，企業・自治体向けの汎用 CMS や，個人向けの SNS・ブログのようなものまで，多種多様なものが存在する．

本研究では，インターネット上でコミュニケーション活動を支援する CMS の中でも，企業・自治体・地域事業といったコミュニティが利用する CMS を想定する．具体的な例として，地域 SNS や企業・事業向けの汎用 CMS が挙げられる．

(1) SNS

SNS とは，人と人とのつながりを促進・サポートする，コミュニティ型のサイトである．友人・知人間のコミュニケーションを円滑にする手段や場を提供したり，趣味や嗜好, 地域, 出身, 友人の友人といった繋がりを通じて新たな人間関係を構築する．

(2) 汎用 CMS

汎用 CMS とは，充実したユーザインターフェイスと拡張性を持つ CMS で，一般サイトからグループウェアまで，多種多様なサイトを構築・運用することができる．

3.3 システムの処理

Step1: プロファイルの作成

コミュニティが利用する CMS に登録されたテキストデータに対して形態素解析 [6] を行い，単語を得る．次に，得られた単語の中から *tf・idf* を用いて，キーワードとなる特徴語 k を抽出する（式 1）．抽出には *tf・idf* を用いる．*tf・idf* とは，テキスト中の単語に着目し，その出現回数や出現範囲などから重要度を設定する手法である．

次に抽出した特徴語 k の関連語を抽出する．関連語は，jaccard 係数 [7] により得られた共起語 kw_i が一定以上のものを抽出して用いる（式 2）．

$$t_{ij} = \log\left(1 + f_{ij}\right) \cdot \log\left(\frac{n}{n_i}\right) \ldots (1)$$

i : 単語のインデックス
j : 文書のインデックス
w_i : 文書集合中における単語
f_{ij} : 単語w_iの文書T_jにおける出現頻度
n : 文書集合中の文書数
n_i : 単語w_iを含む文書数
t_{ij} : 文書T_jに対する単語w_iの重み

関連語であるか否かを判別するための閾値を α としたとき, 関連語である単語 w_i の重み q_i は $kw_i \geq a$ のとき $q_i = 1$，$kw_i < a$ のとき $q_i = 0$ とする．このとき，ある特徴語における関連語の重みを示すベクトルは $q = (q_1, q_2, q_3, \ldots, q_m)$ と表される．特徴語とその関連語ベクトルの集まりをプロファイルとして定義する．

$$kw_i = \frac{\left|k \cap w_i\right|}{\left|k \cup w_i\right|} \ldots (2)$$

w_i : CMS内のテキストの単語
k : t_{ij}の値が高い特徴語となるw_i
kw_i : kとw_iの共起度

Step2: 候補ウェブ記事の取得

ウェブ記事収集の際には，既存の検索エンジンを利用し，対象事業・分野名など，コミュニティを示す単語と特徴語 k を組み合わせて検索し，候補ウェブ記事 $D_1, D_2, \ldots, D_n$ を得る．次に，得られた記事 D_j に対し形態素解析を行い， m 個の単語 $x_1, x_2, \ldots, x_m$ を得る．単語 x_m の文書 D_j における重みを d_{mj} とすると文書 D_j のベクトルは $d_j = (d_{1j}, d_{2j}, d_{3j}, \ldots, d_{mj})$ と表される．

Step3: 類似度によるランキング

内容の類似度に基づいたランキングを，得られた関連語ベクトル q と候補ウェブ記事のベクトルである d_j との比較によって行う. 類似度の算出には式(3)のベクトル空間モデルにおけるコサイン尺度 [8][9] を用いる.プロファイルと候補記事の類似度の値 $\cos(d_j, q)$ が大きいものは，CMS 内での話題に近い記事であり，対象とするコミュニティにとって興味のあるウェブ記事であると判定される．

$$\cos(d_j, q) = \frac{d_j \cdot q}{\|d_j\| \|q\|} = \frac{\sum_{i=1}^{m} d_{ij} q_i}{\sqrt{\sum_{j=1}^{m} {d_{ij}}^2} \sqrt{\sum_{i=1}^{m} {q_i}^2}} \ldots (3)$$

4. 学童保育情報サイトの開発と検証実験

4.1 対象事業の概要

実験では，学童保育コミュニティを対象とした情報サイトである岩手県学童保育情報サイト（以下，情報サイトと呼ぶ）に対して，提案するニュースサイトを実際に導入する．また，提案した収集手法の有効性の確認のため，実際に得られた特徴語を用いてウェブ記事を収集し，類似度によるランキングの評価を行う．

学童保育 [10] とは小学校に通っている学童の放課後の生活を守る事業である．岩手県を例に挙げると，200 箇所以上の学童保育所（児童館などを含む）が存在しており，地区市町村の保育所を統括する地区連絡協議会（以後，地区連協と呼ぶ）と，地区連協を取りまとめる県連絡協議会（以後，県連協と呼ぶ）が存在している．また，学童保育所は公立，公営，社会福祉協議会，運営委員会など，様々な組織によって運営されているため，運営基準も地域によって大きく異なっている．

4.2 岩手県学童保育情報サイトの開発 [11][12]

情報サイト（図 3）は，学童保育事業における情報共有・発信を実現するために構築され，試験運用を経て 2007 年 5 月から正式に運用を開始した．2009 年 7 月時点での参加団体数は，岩手県内の約 260 箇所の学童保育所に，県連協および地区連協を加えた 267 団体となっている．利用対象者は，県連協職員，地区連協職員，指導員，地域住民および入所者・入所希望者の保護者である．情報サイトは，学童保育向けの汎用 CMS として(1)～(4)の特徴を備えている．

表 2　開発環境

	サーバーの環境	クライアント環境
WWW サーバ	Apache2.2.2Tomcat5.5.12	Tomcat5.5.9
DB	MySQL5.0	MySQL5.0
CPU	Intel Pentium D	Intel Celeron M
OS	Linux(Fedora Core5)	Microsoft Windows XP

(1)　コンテンツの収集と作成
　学童保育所の紹介，行事，お知らせなどのコンテンツの登録を行うことができる．一部のコンテンツではテキストだけではなく画像を扱うことも可能となっている．
(2)　ウェブページの構成管理
　CSS (Cascading Style Sheets) を用いることにより，ステータスを変更するだけでホームページのデザイン構成を容易に変更することができる．
(3)　配信管理
　全ての登録情報に関して，公開・非公開の設定を行うことができる．
(4)　利用者管理
　利用者に対して管理者，県・地区連協職員，保育所指導員などロールを設定することで，アクセスの権限の設定を行う．

　また，2008 年に工藤は，KT 法を用いることで議論を活性化させる仕組みを持ったコミュニティサイト [13]を提案し，本情報サイトへ適用した．これにより，利用者同士のコミュニケーションが可能となり，コミュニティサイト上で他利用者からのレスポンスを得られることによって，CMS の利用意欲の向上へと繋がると考えられる．

4.3　実装

　本情報サイトを対象とし，ウェブ記事収集システムを備えたニュースサイトの実装を行った．サーバー及びクライアントの開発環境を表 2 に示す．また，提案システム導入後の情報サイトの構成を図 2 に示す．
　提案するウェブ記事収集システムでは，既存のシステムである学童保育向けの汎用 CMS から抽出された特徴語に基づき，候補となるウェブ記事を収集していく．そのため，既存のシステムをベースとして拡張する形で実装した．また，特徴語の抽出の際には，形態素解析ツールである Sen を利用した．
　システムで収集された候補ウェブ記事は，類似度に基づきランキングされ，運営者に提示される．運営者は提供するに相応しいウェブ記事を選択し，ニュースサイトにて提供する．ニュースサイトを導入し，継続的にニュースを配信することで，既存のコミュニティサイトにおける情報収集面での満足度の向上・利用者同士のコミュニケーションの促進が期待できる．また，コミュニティサイトでは，電子掲示板だけではなく，コメント機能，質問コーナー，テーマ・議題設定機能など，ニュースサイトを通して交流を促すための仕組みが実装されている．拡張実装後の情報サイトの実行画面を図 3 に示す．

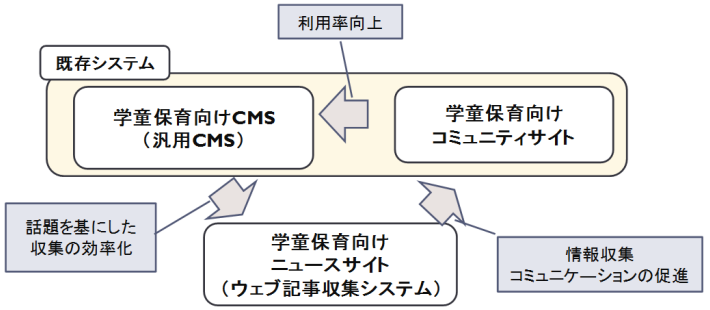

図 2　学童保育情報サイトへの適用

図 3　情報サイトの実行画面

4.4 検証実験

4.4.1 実験用データ

提案した収集手法の有効性の確認のため，本情報サイトのデータを用いた検証実験を行う．まず，学童保育向け CMS に登録されたテキストを基に，学童保育事業従事者の興味を表すプロファイルを作成する．さらに，検索エンジンを用いて学童保育事業に関連する候補記事を実際に収集する．収集した記事に対して適合・不適合の分類を行い，評価用のデータセットとする．

(1) プロファイルの作成

情報サイトでは，お知らせやイベントのテキスト情報がコンテンツとして登録されており，そのコンテンツを 1 つの文書単位と見なした．分析は 1 カ月のテキスト情報を単位として行い，主に季節性のある話題を特徴語として抽出した．各月のコンテンツ数，単語数，具体的な特徴語の例について表 3 に示す．

(2) データセットの作成

本研究の評価実験では，独自のデータセットを作成した．本事業に背景理解のある 3 名で，ニュースとして適性のあるウェブ記事を手作業でウェブ上から収集した．検索エンジン Google にて，事業名と，2008 年 12 月～2009 年 5 月にかけて出現した話題として適切であると思われる特徴語を組み合わせた検索を行い，出力された上位 100 件のウェブ記事を適合・不適合に分類した．このとき，単に事業名と特徴語に文法的な関連があるだけではなく，ニュースとして提供するのに適切であると判断された記事のみを適合文書とした．具体的には，個人的なブログや学童保育所の入所案内・行事表など，限定された利用者にのみ価値のある記事は不適合文書とした．適合文書の記事の例を表 4 に示す．

表 3　抽出された特徴語（2008 年 12 月～2009 年 6 月）

	イベント	お知らせ	総単語数	特徴語
2008 年 12 月	28	13	1773	クリスマス，年末年始，冬休み
2009 年 1 月	35	12	2239	ハンドボール，委託料，
2 月	16	13	574	節分，豆，ポスター，送別会
3 月	24	11	1090	春休み，中学生，リニューアル，卒業
4 月	25	3	699	引越し，手紙，宴会，
5 月	37	17	1025	鯉のぼり，運動会，クマ
6 月	30	11	580	バナナ，エコバック，茅の輪

表 4　正解データセット（適合文書）の例

特徴語	記事
クリスマス（7 件）	学童保育でクリスマス会の幹事になりましたが
	主婦のための情報交換ひろば
	クリスマスの手作り講座
	枡形こども文化センター集会室がクリスマス用に飾り付け
春休み（11 件）	学童保育 入所方法
	小学校【春休みのアルバイト】: アルバイト情報ガイド
	子供の春休みはどうしたら… - 教えて！goo
卒業（7 件）	保育園卒業後，学童保育に行きますか？
	学童保育を卒業したらどうしてます？
運動会（5 件）	幼稚園・保育園の運動会弁当
	学童保育の会親子運動会開会式
引越し(17 件)	学童保育事情教えてください． - 教えて！goo
	学童保育のシステム／ワーキングマザーのなんでも相談室
冬休み（5 件）	忘れないで！学童保育申し込み
鯉のぼり（3 件）	鯉のぼり掲揚式

4.4.2 数値実験

実験では，特徴語と関連のある単語を関連語として判定するための閾値を変化させてランキングの評価を行った．それぞれ閾値が 0.01, 0.03, 0.05, 0.07, 0.1 の場合について検証した．評価にはあらかじめ決められた n 個の再現率レベルでの適合率の平均値である n 点平均適合率を用いた．その際，n は情報検索システムの評価で一般的に用いられる 11 点とした．また，再現率レベルでの適合率を求めるため，再現率を R_i，適合率を P_i とした補完適合率を用いた．補完適合率と 11 点平均適合率はそれぞれ式(4),(5)で表される．各特徴語の平均適合率と関連語数を表 5 に示す．なお，網掛けで示した部分は，各特徴語において最も適合率が高かった値である．

Vol.2009-IS-110 No.6
2009/12/1

表 5　11 点平均適合率（関連語数）

特徴語 閾値 α	ｸﾘｽﾏｽ 7 件	卒業 7 件	春休み 11 件	運動会 5 件	引越し 17 件	冬休み 5 件	鯉のぼり 3 件
0.01	0.23(277)	0.32(181)	0.16(31)	0.11(182)	0.31(72)	0.38(463)	0.06(14)
0.03	0.22(183)	0.12(142)	0.17(21)	0.13(123)	0.34(63)	0.27(107)	0.06(13)
0.05	0.25 (31)	0.10(98)	0.17(17)	0.13(50)	0.36(56)	0.13(34)	0.06(13)
0.07	0.18 (16)	0.09(67)	0.19(14)	0.05(20)	0.43(44)	0.15(8)	0.05(11)
0.1	0.20(5)	0.10(11)	0.20(5)	0.05(10)	0.48(34)	0.53(1)	0.04(10)

$$P(x) = \max_{x \le R_i} P_i \ldots(4) \qquad \tilde{P} = \frac{1}{11}\sum_{i=0}^{10} P\left(\frac{i}{10}\right) \ldots(5)$$

4.5 考察

提案したシステムにより，学童保育に関連しないウェブ記事，プライベートなウェブ記事といったものを除外することができ，一部の特徴語においては大幅に精度が改善された．全ての特徴語においても，50 位以降の適合文書が 20～40 位程度に上昇する傾向が見られた．

しかし，プロファイルが不適合文書に多く含まれる単語を含むケースや，プロファイルが適合文書において重要な単語を含まないケースが見受けられた．

具体的な例として，特徴語「卒業」において，関連語「教育」「学校」といった単語により，大学の教育機関や学歴のページといった不適合文書の類似度が高くなった．また，特徴語「春休み」においては，アルバイトに関する適合文書が多く見受けられたが，アルバイトを示す単語はプロファイルには含まれていなかったため，類似度が高くはならなかった．

これらから，プロファイルとの類似度を指標とするだけでは，類似度が高いがニュース性が低いケースや，類似度が低いがニュース性が高いといったウェブ記事の分類が難しいという結果が得られた．

5. 適合性フィードバックによるプロファイル更新手法

5.1 目的

実験結果から，プロファイルとの内容の類似度だけでは，類似度の高い不適合文書や類似度の低い適合文書の分類は難しいといった結果が得られた．そこで，検索結果を基に，適合文書・不適合文書の特徴を学習するため，CMS プロファイルを適合性フィードバックに基づき更新する手法[14]を提案する．

5.2 適合性フィードバックの適用

適合性フィードバックとは，利用者が得られた検索結果のうち，どの文書が検索意図に適合し，どの文書が適合しないかを検索システムに学習させることにより，検索精度を改善する手法である．適合性フィードバックでは，適合文書に含まれる単語の重みを大きく，不適合文書に含まれる単語の重みを小さくし，検索質問中の単語の重みを調整する．

$$q' = \alpha q + \frac{\beta}{|D_R|}\sum_{d_i \in D_R} d_i - \frac{\gamma}{|D_N|}\sum_{d_j \in D_N} d_j \ldots(6)$$

α : 元のプロファイルをどれだけ重視するか
β : 適合文書中の単語をどれだけ重視するか
γ : 不適合文書中の単語をどれだけ重視するか
D_R :フィードバックする適合文書数
D_N :フィードバックする不適合文書数

本研究では，適合文書および不適合文書による重みの調整分を，それぞれの文書集合に含まれる文書数で正規化したロッチオの式を用いる．

検索結果の中に含まれる適合文書の集合を D_R 不適合文書の集合を D_N とした場合，式(6)を用いて検索質問ベクトルであるプロファイル q を q' に修正する．式(6)によって得られた単語のうち，重みの高い単語のベクトルを 1 に，重みの低い単語のベクトルを 0 に正規化し，プロファイルを更新する．なお，α, β, γ は 0 以上の定数であり，それぞれ検索質問であるプロファイル，適合文書，不適合文書に出現する単語をどの程度重要視するかを示している．また，類似度の低い適合文書に関しては，運営者が適切な記事を任意に加えることでプロファイルを更新する．

これにより，適合文書に含まれる単語を追加・不適合文書に含まれる単語を除去し，プロファイルを更新することで，適合率の向上を図り，収集の効率化を実現する．

5.3 数値実験

実験では，任意の定数である α, β, γ の値と，フィードバックする単語数の組み合わせによって式(6)から算出された単語の情報を基にプロファイルを更新し，類似度を再算出する．評価には，前実験と同様，式(4)の補完適合率と式(5)の 11 点平均適合率を用いる．任意の定数には，SMART (TREC-7) [15]の値設定である α =3, β =2, γ =2 と，CAFES (TREC-8)[16]の値設定である α =1, β =0.1, γ =0 を用いた．また，フィードバックを行う単語をそれぞれ変化させて行った．実験パターンをⅠ～Ⅳとし，それ

表 6　プロファイル更新後の 11 点平均適合率（関連語数）

特徴語 閾値 α	ｸﾘｽﾏｽ 7 件	卒業 7 件	春休み 11 件	運動会 5 件	引越し 17 件	冬休み 5 件	鯉のぼり 3 件
前実験	0.25(31)	0.32(181)	0.20(5)	0.13(50)	0.48(34)	0.53(1)	0.06(13)
Ⅰ	0.27(36)	0.39(184)	0.29 (9)	0.19(52)	0.64(41)	0.72(4)	0.66(20)
Ⅱ	0.36(38)	0.63(188)	0.57(17)	0.25(56)	0.64(41)	0.72(7)	0.82(30)
Ⅲ	0.25(31)	0.32(181)	0.20(5)	0.13(50)	0.57 (38)	0.53(1)	0.44(17)
Ⅳ	0.25(31)	0.32(181)	0.46 (27)	0.13(50)	0.57 (38)	0.53(1)	0.75(30)

ぞれ前の実験にて高い結果が出た各特徴語のプロファイルに対して，単語をフィードバックし，類似度を再算出した．

Ⅰ．SMART(TREC-7) ：$\alpha=3$，$\beta=2$，$\gamma=2$　　**フィードバックバック単語数 10**
Ⅱ．SMART(TREC-7) ：$\alpha=3$，$\beta=2$，$\gamma=2$　　**フィードバックバック単語数 20**
Ⅲ．CAFES(TREC-8) ：$\alpha=1$，$\beta=0.1$，$\gamma=0$　　**フィードバックバック単語数 10**
Ⅳ．CAFES(TREC-8) ：$\alpha=1$，$\beta=0.1$，$\gamma=0$　　**フィードバックバック単語数 20**

ⅠとⅡは，情報検索システム SMART で用いられた値設定であり，適合文書と不適合文書を同程度重視するものである．Ⅲ，Ⅳは CAFES による値設定であり，プロファイルを重視し，適合文書中の単語のみを考慮するものである．フィードバックする文書数は，適合文書数と同じ数だけの不適合文書とする．Ⅰ,Ⅱでは値の高い単語・値の低い単語をそれぞれ 10, 20 と変化させ，プロファイル対して追加・除去する処理を行った．Ⅲ，Ⅳでは，適合文書の特徴のみを考慮しているため，重みの高い単語のみを 10, 20 と変化させてプロファイルに追加を行った．その結果を表 6 に示す．

5.4 考察

結果として，プロファイルを更新する手法を取り入れたことで精度が改善され，特に実験条件のⅡでは，全ての特徴語において大幅な適合率の向上が得られた．

Ⅲ，Ⅳにおいては，α に比べ β は低く，不適合文書をまったく考慮しなかったため，もともとプロファイルに含まれていた単語のみが抽出された．本システムでは，膨大なウェブ記事の中からニュース性のある記事のみを収集するという目的からも，適合文書に対して不適合文書の割合が高い傾向をもつため，不適合文書の情報をフィードバックする必要があると考えられる．

図 4　実行画面（フィードバックによる更新前）

図 5　実行画面（フィードバックによる更新後）

各特徴語に関しては，「クリスマス」「引越し」など，元の結果が良かった特徴語よりも，「卒業」「春休み」といった元の結果が良くなかった特徴語の精度が向上した．具体的には，特徴後「卒業」においては，大学の教育機関や学歴などのページに含まれる「学校」「教育」「学生」といった単語を除去することができた．特徴語「春休み」においては，学童保育のアルバイトに関する Q&A の記事に頻繁に含まれる「募集」「質問」「回答」といった単語を抽出することができた．

システムの実行例として，内容の類似度だけを考慮した結果を図 5 に，フィードバックによるプロファイル更新を行った結果を図 6 に示す．プロファイルの更新を行うことで，提供すべきウェブ記事が上位にランクインしていることが分かる．この仕組みにより，今後は膨大な記事にも対応でき，運営者がより効率的に提供するべき記事を選択することが可能となった．今後は，任意の値を細かに変化させて数値実験を行い，より適切な値設定を求め，精度の向上を図る．

6. おわりに

本研究では，特定のコミュニティを支援する CMS の活性化を図る一方策として，CMS の話題に基づいたウェブ記事を収集するニュースサイトを構築した．収集の際には，ベクトル空間モデルに基づき，CMS 内のプロファイルとの類似度に基づきランキングで出力することで，運営者が提供すべき記事の選択の支援を実現した．さらに，内容の類似度だけではなく，CMS 内の話題を示すプロファイルとの類似度の低い適合文書や高い不適合文書といった文書の特徴を考慮するため，適合性フィードバックを取り入れたプロファイルの更新手法を提案した．これにより，適合文書の見落としや不適合文書の誤選択を防ぐことができ，さらに，候補となるウェブ記事が膨大となる場合でも，上位の記事に対して判定するだけで適合文書を容易に取得することが可能となった．

今後は，実験で対象とした学童保育コミュニティにて引き続き数値実験を行う．具体的には，今後の実験以後の運用で得られた特徴語を基にしてプロファイルを作成し，よりフィードバックの値を細かに変化させた場合や候補となるウェブ記事を増加させた場合についても検証する．さらに，学童保育コミュニティを対象として，ウェブ記事を継続的に収集・配信し，ニュースサイトを導入することによるサイト活性化の観点からの評価を行う．

参考文献

1) 小笠原盛浩：オンラインコミュニティ類型を用いた利用と満足分析-日韓学生データを用いた利用行動の探索的研究-, The Japan Association for Social Informations, (2006)
2) 鈴木康祐, 岡本東, 堀川三好, 菅原光政：特定分野・事業の文書集合を利用したニュース記事収集システムの提案, 情報処理学会第 71 回全国大会講演論文集, 分冊 4, pp.487-488, (2008)
3) newsing：http://newsing.jp/,（最終アクセス日：2009/010/3）
4) Buzzurl：http://buzzurl.jp/,（最終アクセス日：2009/10/3）
5) Digg：http://digg.com/,（最終アクセス日：2009/10/3）
6) Sen とは：http://ultimania.org/sen/,（最終アクセス日：2008/01/20）
7) 相澤彰子：大規模テキストコーパスを用いた語の類似度計算に関する考察, 情報処理学会論文誌, Vol.49, No.3, pp.1426-1436, (2008)
8) 北研二, 津田和彦, 獅々堀正幹：情報検索アルゴリズム, 共立出版株式会社, (2008)
9) 徳永健伸：情報検索と言語処理, 東京大学出版会, (1999)
10) 全国学童保育連絡協議会：http://www2s.biglobe.ne.jp/~Gakudou/,（最終アクセス日：2009/06/21）
11) 舘澤千尋, 岡本東, 堀川三好, 菅原光政：学童保育を対象としたコンテンツ管理システム, 情報処理学会第 69 回全国大会講演論文集, 分冊 4, pp.151-152, (2007)
12) 伊吹山香理, 岡本東, 堀川三好, 菅原光政：学童保育事業における事務支援システム, 情報処理学会第 68 回全国大会講演論文集, 分冊 4, pp.591-592, (2006)
13) 工藤拓也：グループ問題解決支援システムの開発, 岩手県立大学ソフトウェア情報学部卒業論文, (2008)
14) 帆足啓一郎他：非適合プロファイルを利用した文書フィルタリング手法, 情報処理学会論文誌, Vol.42, No.3, pp.507-517, (2001)
15) Amit Singhal,John Choi,Donald Hindle,David D.Lewis and Fernand Pereira：AT&T at TREC-7, The 7th Text Retrieval Conference, pp.239-251, (1999)
16) Chengxiang Zhai,Peter Jansen,Norbert Roma,Emilia Stoica and David A.Evans：Optimization in CLARIT TREC-8 Adaptive Filtering, 8th Text Retrieval Conference, pp.253-258, (2000)