

Wiki などからの組織プロフィールとクローラを利用した 情報収集

徳永秀和[†] 片岡啓介^{††}

近年の Blog など個人が簡単に Web ページを作成することが可能となり、Web に存在する情報は爆発的に増えている。そのため、Web 全体を把握することは極めて困難になってきており、Google など現在の汎用的な検索エンジンでは、ユーザが自身にとって有益である情報を追いつけるのは困難であり、ユーザの嗜好や状況に合った Web ページをフィルタリングし、推薦するユーザ適応型の情報推薦システムの重要性が高まっている。この情報フィルタリング手法として協調フィルタリングと内容ベースフィルタリングという 2 つのアプローチがある。多くのユーザとは異なる興味を満たす Web ページを推薦するには内容ベースフィルタリングが有効である。

本論文では、グループで情報を共有し発信する Wiki や Blog を対象にし、組織プロフィールを作成する。そして、組織プロフィールをもとにクローリングをすることにより、グループの興味にあった Web ページを収集、推薦するシステム開発について説明し、その評価実験結果を報告する。

Web Page Recommender System Using a Group Profile

Hidekazu Tokunaga[†] and Keisuke Kataoka^{††}

In this paper, the Web page recommender system using crawler and the group profile which were created using Web Services, such as Blog, is proposed. The system carries out the automatic collection and recommendation of a Web page which suited the profile using contents base filtering of a vector space model. The profile was expressed as a vector some documents of the Web Service which the user uses. The crawler performs filtering processing and crawling simultaneously and collects pages which suited the profile. The effectiveness of the system is evaluated by experiments. Subjects are three profiles. As a result, it is shown that the information filtering using a profile was possible. In addition it is shown that collection of pages using the crawler was also possible. However, in a certain profile including the contents of various fields, the system did not function well. Since the present system is not considering a contribution day of a document, if these are improved, recommendation of the page which suited a user more is possible.

1. はじめに

近年の高度情報化にともない Web に存在する情報は爆発的に増えている。Google の発表によると、1998 年に初めてインデックスを作成した時点で既に 2600 万ページを数えたが、2008 年には同社が把握した Web ページの数は 1 兆ページを突破し、今現在も 1 日に数 10 億ページが新たに生成されているという[1]。そのため、Web 全体を把握することは極めて困難になってきており、Google など現在の汎用的な検索エンジンのように、Web ページの増加に追従して収集やインデックス化を行い、かつユーザに対する検索結果の精度を維持するには、設備投資や情報鮮度の低下などの点で限界がある。このような状況の中でユーザが自身にとって有益である情報を汎用的な検索エンジンのみで追いつけるのは困難であり、ユーザの嗜好や状況に合った Web ページを推薦するユーザ適応型の情報推薦システムの重要性は高いと思われる。

ユーザの興味や状況に適応した情報を推薦するには、未知の情報をユーザの嗜好を考慮してフィルタリングする必要がある。この情報フィルタリング手法として「協調フィルタリング」と「内容ベースフィルタリング」という 2 つのアプローチがある。「協調フィルタリング」とは、多くのユーザの嗜好情報を蓄積し、あるユーザと嗜好の類似した他のユーザの情報を用いて自動的に推薦を行う手法である。「内容ベースフィルタリング」とは、文書の内容とユーザの嗜好や関心を表すユーザプロフィールとをマッチングすることで、ユーザの嗜好に合った文書を選別する手法である。「内容ベースフィルタリング」には、推薦元となる Web ページをシステムやユーザ自身が探さなければならない、ユーザの新たな嗜好を得るためにはユーザプロフィールの更新を行わなければならないといった問題点があり、ユーザ適応型のシステムにおいては「協調フィルタリング」に比べ利用機会は少ない。しかし、「協調フィルタリング」では処理できない、ユーザの誰も評価を与えていない情報、例えば他ユーザからの推薦が得られにくいニッチな情報や作られたばかりの情報などは「内容ベースフィルタリング」であれば情報フィルタリングが可能となる。

本論文では、組織において知識共有の効率化を目的として利用されている Blog や wiki, SNS などの Web サービスを対象として、組織が持っている興味を組織プロフィールとして作成する。そして、Yahoo などの汎用的な検索サイトとクローラを用いた「内容ベースフィルタリング」による Web ページ推薦システムの開発およびその有効性の検証実験の結果を報告する。

[†] 香川高等専門学校機械電子工学科

Department of Electoro-Mechanical Systems Engineering, Kagawa National College of Technology

^{††} 香川高等専門学校創造工学専攻

Advanced Course in Industrial Systems Engineering, Kagawa National College of Technology

2. 提案システム

2.1 提案システムの構成

提案システムではベクトル空間モデルを用いた内容ベースフィルタリングを用いて、プロフィールに適合した Web ページの自動収集および推薦を行う。提案システムの構成図を図 1 に示す。

まず、組織で利用している Blog, wiki, SNS などから組織プロフィールの作成を行う (図中①)。この組織プロフィールから重要単語をいくつか抽出し、既存検索エンジンで重要単語に対する検索結果の上位 URL を取得し、「スタート URL」として記憶する (図中②)。全ての「スタート URL」中の URL に対してクローラをばらまき、各クローラはページの情報フィルタリングを行う (図中③)。情報フィルタリングは、内容ベースフィルタリングを用いて組織プロフィールと Web ページとの間の類似度を算出して行う。類似度が閾値以下であれば組織プロフィールに適合していない Web ページであると判断し、URL、タイトルおよび類似度を保存した上でクローラは消滅し、それ以上リンクはたどらない (図中④)。類似度が閾値以上であれば適応した Web ページであると判断し、URL、タイトル、類似度、そのページからのリンクを保存した後クローラは増殖する (図中⑤)。増殖したクローラは更に保存したリンクをたどり同様の処理を行う (図中⑥)。

このようにスタート URL へばらまいたクローラが増殖もしくは消滅を繰り返すことで、ユーザに適応した Web ページの収集を行うことができる。収集した Web ページは類似度に基づいたランク付けを行った上でユーザへ提示する (図中⑦)。

2.2 開発環境

開発環境を表 1 に示す。システム開発はロボット型検索エンジン"SUZAKU"[2]をベースとして行った。

2.3 情報フィルタリング方法

ユーザの嗜好に適合した情報を選別する情報フィルタリング方式には、内容ベースフィルタリングと協調フィルタリングの 2通りのアプローチがある。協調フィルタリングは嗜好の類似した他のユーザからの推薦を自動化したものである。内容ベースフィルタリングは文書の内容とユーザプロフィールとをマッチングすることで、ユーザの嗜好に適合した文書を選別する方式である。本システムでは、出現単語に基づいて文書を 1つのベクトルとして表現し、ベクトルの向きによって内容を判断するベクトル空間モデル[3]を適用して内容ベースフィルタリングを実現する。

2.3.1 組織プロフィール作成

ここでは、組織プロフィール作成 (図 1-①) について説明する。Blog や wiki, SNS などの Web サービスで書かれた文書は、一般に利用しているユーザの興味や嗜好を表していると考えられる。本システムでは、wiki などの Web サービスで管理される複数

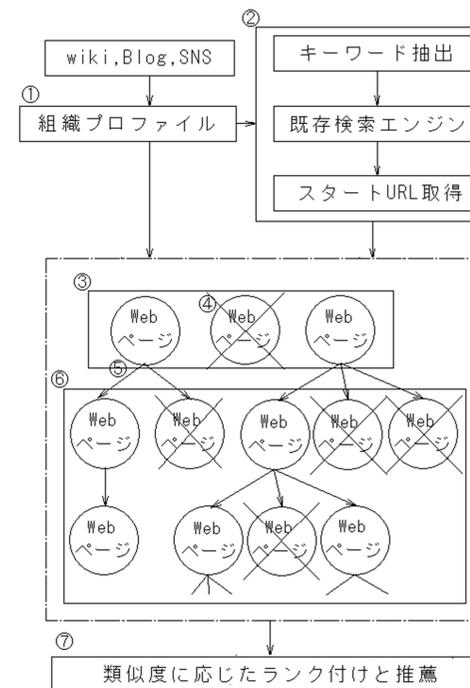


図 1 システム構成図

表 1 開発環境

用途	使用ソフト
OS	Ubuntu 9.10
開発言語	Ruby (Ruby 1.8.6)
Web サーバ	Apache 2.2.3
データベースソフト	MySQL 5.0.27
形態素解析ソフト	Chasen 2.3.3

の文書をベクトル空間モデルにより 1つのベクトルで表現したものを組織プロフィールとして扱う。

組織プロファイルの作成手順について記述する。ユーザによって与えられた、Web サービスで管理される複数の文書を1つの文書 Q と考える。文書 Q を形態素解析ソフト「茶筌」[4]を用いて文書解析を行い有効語の候補を抽出する。一般には文書中の単語のうち重要度の高いと思われる名詞や形容動詞の語幹、サ変動詞の語幹が有効語の候補とされる[5]ため、それらを有効語の候補として重み付けを行う。文書 Q における単語 T の重み $w(T, Q)$ は $tf \cdot idf$ 法を用いて以下のように定義される。

$$w(T, Q) = tf(T, Q) \cdot idf(T) \cdot \dots \cdot (2-1)$$

$tf(T, Q)$, $idf(T)$ はそれぞれ文書に対する単語の出現頻度 tf と逆文書頻度 idf を表した値である。文書 Q における単語 T の出現頻度を $tfreq(T, Q)$ とすると、 $tf(T, Q)$, $idf(T)$ はそれぞれ以下のように計算する。

$$tf(T, Q) = tfreq(T, Q) = \left(\frac{\text{文書}Q\text{中の単語}T\text{の出現回数}}{\text{文書}Q\text{の総単語数}} \right),$$

$$idf(T) = \log_{10} \left(\frac{\text{全文書数}}{\text{単語}T\text{を含む文書数}} \right) + 1.$$

本システムでは「単語 T を含む文書数」を検索エンジン"Yahoo!"での単語 T の検索ヒット数、「全文書数」を"Yahoo!"で検索できる全ての Web ページ数として $idf(T)$ を計算する。"Yahoo!"での単語 T の検索ヒット数は"Yahoo!検索 Web API"を用いて取得する。"Yahoo!"で検索できる全ての Web ページ数は、2005年に"Yahoo! Inc."がインデックス数 192 億ページと発表しているが[6], その後の程度の Web ページ数をインデックス化しているか不明である。インデックス化している Web ページ数は、どの単語においても共通の値をとり各単語の逆文書頻度 idf への相対的な影響はない。オーダーレベルの誤差でなければ単語の重み付けには影響がないとして、200 億ページと仮定して計算を行う。出現頻度 tf は文書長の影響を受けやすいため、以下のような正規化を行った式を用いる。ここで、 $mum(Q)$ を文書 Q に含まれる単語数とする。

$$tf(T, Q) = \left(\frac{\log_{10}(tfreq(T, Q) + 1)}{\log_{10}(mum(Q))} \right).$$

全ての有効語の候補に対して重み付けを行い、重みの大きな単語の上位 m 個を有効語 $V_1, V_2, V_3, \dots, V_m$, としたとき、文書 Q を以下のような m 次元ベクトル \vec{q} として表現したものを組織プロファイルとする。

$$\vec{q} = (w(V_1, Q), w(V_2, Q), \dots, w(V_m, Q))$$

keyword	idf	weight	count
情報	18.4671	0.00140862	14
ユーザ	8.22247	0.000806295	18
フィルタ	8.95188	0.000585309	12
リング	8.28609	0.000541777	12
嗜好	9.76886	0.00037264	7
推薦	8.29342	0.000316358	7
検索	18.4885	0.000302286	3
内容	8.79449	0.000287555	6
協調	11.0566	0.000241027	4
ページ	8.22626	0.000224425	5

図 2 組織プロファイル

図 3 組織プロファイル作成画面

例として第 1 章の本文を用いたプロフィール作成結果を図 2 に示す。図中の項目 idf は逆文書頻度 $idf(T)$, $weight$ は重み $w(T, Q)$, $count$ は文書中における各単語の出現回数である。

プロフィール作成画面を図 3 に示す。ユーザは、自身が利用している Web サービスの URL を入力し「プロフィール作成」ボタンを押すことでプロフィールの作成を行えるようにした。ユーザはリンク探索の深さを 0 から 5 までの整数と「サイト内の全文書」から選択することが可能である。0 なら入力した URL のみをプロフィール作成の対象とし、1 以上ならその値分リンクをたどり、たどった全ての文書をプロフィールの対象とする。「サイト内の全文書」なら同一サイト内全ての文書をプロフィールの対象とする。

2.3.2 収集ページの文書ベクトル作成

クローリング中にクローラが発見した文書を D_1, D_2, \dots, D_n としたとき文書 D_j を以下の

ようなベクトルで表現する.

$$\vec{d}_j = w(V_1, D_j), w(V_2, D_j), \dots, w(V_i, D_j), \dots, w(V_m, D_j)$$

ここで, $w(V_i, D_j)$ は文書 D_j に対する組織プロフィールの有効語 V_i の重みであり, 前項と同様に(2.1)式で計算される.

2.3.3 類似度の計算

組織プロフィールと収集ページの文書ベクトルの類似度 $\text{sim}(Q, D_j)$ は, 各文書を表す文書ベクトル \vec{d}_j と \vec{q} のなす角の余弦値により求められ,

$$\text{sim}(Q, D_j) = \frac{\vec{d}_j \cdot \vec{q}}{\|\vec{d}_j\| \cdot \|\vec{q}\|} = \frac{\sum_{i=1}^m w(V_i, D_j)w(V_i, Q)}{\sqrt{\sum_{i=1}^m w(V_i, D_j)^2} \sqrt{\sum_{i=1}^m w(V_i, Q)^2}}$$

で表される. 類似度は 0 以上 1 以下の実数値となり, 値が大きいくほど文書 D_j は組織プロフィールに適合した Web ページであるといえる.

2.4 クロール方法

クローラの初期配置である「スタート URL」(図 2-②)は, "Yahoo!検索 Web API"を用いていくつかの重要単語を検索し, 検索結果の上位ページを「スタート URL」とする. 検索に用いる重要単語は 2.3 節の a)組織プロフィール作成で抽出した有効語のうち重み が最も大きな単語から順に使用する.

クローラ起動画面を図 4 に示す. ユーザが類似度の閾値, 検索に用いる重要単語数, 各単語の検索結果の取得ページ数の設定が行える. 「クロール開始」ボタンで「スタート URL」を取得しクロールを開始する(図 1-②~⑥). 現在, pdf ファイルや画像ファイルなどには対応しておらず, 収集ページは HTML ファイルのみとなっている. 「プロフィール作成 & クロール開始」ボタンではプロフィール作成を行った上でクロールを開始する. その場合のプロフィール作成の設定は前回作成時と同様のものになるようにしている.

2.5 Web ページ推薦画面

Web ページ推薦(図 1-⑦)画面を図 5 に示す. 結果は類似度の高いページから順に表示する.

3. 実験

3.1 実験方法

第 1 章の本文を用いてプロフィールを作成し, 提案システムの設定パラメータの変



図 4 クローラ起動画面

Webページ表示 969 件 [検索時間 0.01 sec]

1. [協調フィルタリング - Wikipedia](http://ja.wikipedia.org/wiki/%E5%8D%94%E8%AA%BF%E3%8B%82%BF%E3%83%AA%E3%83%B3%E3%82%B0) 類似度: 0.771769 [795] <http://ja.wikipedia.org/wiki/%E5%8D%94%E8%AA%BF%E3%8B%82%BF%E3%83%AA%E3%83%B3%E3%82%B0>
2. [協調フィルタリングによるリコメンデーション](http://www.jpo.go.jp/shiryuu/s_sonota/hyujun_gijutsu/net_k) 類似度: 0.7! [832] http://www.jpo.go.jp/shiryuu/s_sonota/hyujun_gijutsu/net_k
3. [協調フィルタリング - @IT情報マネジメント用語事典](http://www.atmarkit.co.jp/aig/04biz/cf.html) 昇 [802] <http://www.atmarkit.co.jp/aig/04biz/cf.html>
4. [TokyoWebmining#8 協調フィルタリングにおける希薄walk](http://www.slideshare.net/komiyatsushi/tokyo-webmining-8cl) 類似度: 0.690991 [798] <http://www.slideshare.net/komiyatsushi/tokyo-webmining-8cl>
5. [協調フィルタリングとは \(Collaborative Filtering\) き, 用語辞典バイナリ](http://www.sophia-it.com/content/%E5%8D%94%E8%AA%BF%E3%8B%82%BF%E3%83%AA%E3%83%B3%E3%82%B0) 類似度: 0.679255 [819] <http://www.sophia-it.com/content/%E5%8D%94%E8%AA%BF%E3%8B%82%BF%E3%83%AA%E3%83%B3%E3%82%B0>

図 5 Web ページ推薦画面

化により推薦 Web ページがどのように変化するかを調べ, プロフィールに対する設定パラメータの指標作成を行う. 指標パラメータを用いて実験を行い, 本システムがユーザに適合した Web ページ推薦が行えているかを評価する. パラメータの指標の汎用性を調査するため, 異なるプロフィールを用いた実験を複数回行った.

本システムにおける評価項目は, 類似度ごとの推薦ページのプロフィールとの適合率, クローラがリンクをたどって獲得した Web ページ数である. プロフィールとの適合率は, 情報フィルタリングが正しく行えているか, また閾値の設定はいくらが妥当であるかを検証するのに用いる. クローラがリンクをたどって獲得した Web ページ数は, クローラによる Web ページの収集が有効であるかを検証するのに用いる.

表2 指標作成実験の結果

類似度	収集ページ数	関連ページ数	適合率 (%)
0.7 以上	3	3	100
0.6-0.7	6	4	66.7
0.5-0.6	116	1	0.862
0.4-0.5	157	3	1.91
0.3-0.4	143	2	1.40
0.3 未満	510	1	0.00196

3.2 パラメータの指標作成

類似度の閾値はどの程度が妥当かを調べるため、推薦 Web ページの類似度 0.1 きざみごとの組織プロフィールとの適合率を調べる実験を行った。組織プロフィールは第 1 章の本文を用いて行った。パラメータの設定は、検索エンジンへの検索単語数を 10、各検索単語の検索エンジンからの URL 取得数を 100 とした。検索単語数は図 2 の組織プロフィール作成結果より、重み付けの上位 10 単語程度が妥当であると判断した。URL 取得数は、適合率算出のための十分な数の収集ページが得られ、かつ筆者が全ての収集ページに対して適合しているかのチェックを行える範囲内かを考慮して 100 とした。類似度の閾値は最高値の 1 とし、クローラが類似度算出後にリンクをたどらないようにした。この時の収集ページは「スタート URL」とほぼ同じ数となるため、

$$(\text{検索単語数 } 10) \times (\text{取得 URL 数 } 100) = 1000 \text{ ページ}$$

となる。

以上の条件で実験を行った結果を表 2 に示す。クローラが訪れて収集した全ページ数である「収集ページ数」、収集ページの中から本論文と関連した Web ページの数である「関連ページ数」、関連ページ数の収集ページ数における割合である「適合率」を類似度別にそれぞれ示している。Web ページが関連ページであるかの判断は筆者が Web ページのタイトルもしくは Web ページ本文を閲覧して行った。

表 2 の実験結果より、類似度が 0.7 以上の場合の適合率が 100%、類似度が 0.6 から 0.7 の場合の適合率 66.7% となっており、類似度が 0.6 以上の Web ページでは適合率 77.8% と高い確率で関連ページを推薦できていることが分かる。類似度が 0.6 より低い Web ページでは、類似度 0.5 から 0.6 で 0.86%、類似度 0.4 から 0.5 で 3.11% と適合率は低い。類似度が 0.6 より小さいページの中にも関連ページは存在しているが、収集ページ数が多く適合率は低い。クローラがプロフィールに適合していないページのリ

表3 Blog の実験結果

類似度	実験a			実験b			実験c		
	収集ページ数	関連ページ数	適合率(%)	収集ページ数	関連ページ数	適合率(%)	収集ページ数	関連ページ数	適合率(%)
0.7以上	64(56)	53(49)	82.8	71(60)	65(58)	91.5	0	0	—
0.6-0.7	19(13)	17(12)	89.4	200(185)	183(178)	91.5	0	0	—
0.5-0.6	29(5)	11(2)	37.9	101(72)	71(62)	70.2	0	0	—
0.4-0.5	86(17)	20(6)	23.2	121(70)	71(58)	58.6	9(0)	6	66.6
0.3-0.4	150(25)	14(2)	9.33	86(33)	47(18)	54.6	81(0)	41	50.6
0.3未満	658(31)	25(3)	3.79	779(159)	72(13)	9.24	728(0)	126	17.3
合計	1006(147)	140(74)	13.9(50.3)	1358(579)	496(387)	36.5(66.8)	818(0)	173(0)	21.1

表4 Wiki の実験結果

類似度	実験d			実験e		
	収集ページ数	関連ページ数	適合率 (%)	収集ページ数	関連ページ数	適合率 (%)
0.7以上	2(1)	1(1)	50	11(0)	10(0)	90.9
0.6-0.7	11(1)	8(1)	72.7	29(4)	22(3)	75.8
0.5-0.6	13(3)	6(2)	46.1	23(5)	17(3)	73.9
0.4-0.5	17(0)	6(0)	35.3	25(4)	12(1)	48
0.3-0.4	50(10)	27(0)	18	25(0)	8(0)	32
0.3未満	845(48)	9(5)	3.19	681(51)	51(13)	2.13
合計	928(63)	57(9)	6.14(14.2)	794(64)	120(20)	17.3(31.2)

リンクをたどるよう設定すると、本システムが把握する Web 空間の、プロフィールとは関係のない空間が増大して効率が悪くなる。そのため、類似度が閾値以上のページの適合率がある程度高くなるように閾値を設定する必要がある。本実験においては類似度の閾値は 0.6 以上程度と考えられる。最適な閾値はプロフィールやシステムを使用するユーザによって変化すると考えられるため、閾値の汎用性についても実験で調査する。

3.3 実験結果

3.2 節の指標作成結果より、実験で使用するパラメータを表 2 のように決定した。3.2 節より類似度の閾値は 0.6 とした。検索エンジンへの検索単語数および各検索単語の検索エンジンからの URL 取得数は 3.2 節における実験パラメータと同様な理由で、組織プロフィールおよび適合率調査を考慮して決定した。表 2 のパラメータを用いて実験を行った結果を表 3 と表 4 に示す。表中の()内の数字は「スタート URL」に登録されている URL を除いた数字で、クローラがリンクをたどって獲得した Web ページの数と等しい。表 3 は、作成者の異なる 3 つの Blog の最新の記事から 20 件を対象にして組織プロフィールを作成し、Web ページ収集を行った。表 4 は、2 つの Wiki に対してトップページよりリンクの深さを 3 として組織プロフィールを作成し、Web ページ収集を行った。

4. 考察

実験 a および実験 b においては、両実験ともに類似度 0.6 以上で適合率 80% 以上となっており、高い確率で関連ページを推薦できている。実験 a に比べ実験 b では類似度 0.6 未満でも適合率が高い。これは、収集ページに本システムでは解析できない動画や画像がメインのページが多く存在するためである。実験 b のような場合には閾値を小さくしたほうが良い結果を得られる可能性がある。提案システムの特徴としてクローラを用いた推薦元となる Web ページの収集があげられるが、実験 a および実験 b のクローラがリンクをたどって得た Web ページ数はそれぞれ 149 ページ、579 ページとなっており、クローラによる Web ページ収集は有効である。

実験 c においては収集ページに類似度が閾値 0.6 以上のページがなく、クローラがリンクをたどって収集したページは 0 であった。しかし、関連ページ数は合計で 173 ページ存在し類似度 0.4-0.6 における適合率は 66.6% となっており、他の実験に比べ Web ページの類似度は小さくなる傾向にある。実験 c の組織プロフィール作成に用いたブログは他の実験に用いたブログに比べると取り扱う分野が広いことが原因である。このようなブログに対しては、文書クラスタリングによりいくつかの類似した文書群に分類した上で、それぞれの文書群に対しプロフィール作成および Web ページ推薦を行う必要がある。

2 つの Wiki については、実験 a および実験 b と同様の傾向にある。これは、Wiki は比較的まとまった内容のページが集まっているためだと考えられる。

5. おわりに

本論文において、Blog などの Web サービスを用いて作成した組織プロフィールとクローラを用いた Web ページ推薦システムの提案、および提案システムの有効性を検証

するための評価実験を行った。

まず現在の Web についての現状を述べ、情報推薦システムの重要性と情報フィルタリング手法について述べた。さらに本システムで利用した内容ベースフィルタリングについての欠点および利点を述べた。内容ベースフィルタリングには、誰も評価を与えていない情報のフィルタリングが可能であるが、推薦元となる情報源を自らで準備しなければならない、ユーザの新たな嗜好を得るためにはプロフィールの更新を行わなければならないといった問題点があった。

問題に対して提案システムの詳細を述べ評価実験を行った。実験 a および実験 b を行った結果、類似度が 0.6 以上では適合率 80% 以上という結果が得られ、組織プロフィールを用いた情報フィルタリングは有効であることが分かった。本システムでは解析できない動画や画像中心の Web ページに対しては有効な情報フィルタリングが行えない問題があり、最適な類似度の閾値はプロフィールなど状況によって変化することが分かった。実験 c においては、様々な分野の記事をプロフィールの対象としたため、他の実験に比べ Web ページの類似度が小さくなる傾向にあった。

今後の課題として、文書クラスタリング機能の追加が挙げられる。また類似度の閾値の決定方法にも課題があり、システムを利用するユーザが自由に類似度の決定を行えるようにするなどの機能が必要である。以前の実験で複合語の取り扱いができないという問題があり、今回は手動で辞書への複合語登録を行ったが、自動で複合語の処理が行えるようになるのが望ましく、対象ページ内の語句の出現順序も考慮して正しい検索結果を求めるなどの機能追加が必要である。提案システムの特徴として Web サービスを用いた組織プロフィールの作成があるが、現在は投稿日が異なる文書でも同じ扱いをしている。ユーザの現在の状況により近い組織プロフィールを正確に得るには、投稿日を考慮した組織プロフィール作成が必要であると思われる。現在はこのような課題を解決できるようにシステムの改良を行っている。

参考文献

- 1) Jesse Alpert & Nissan Hajaj, We knew the web was big.
<http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html>
- 2) 星澤隆: Ruby でつくる検索エンジン, 株式会社毎日コミュニケーションズ, (2006)
- 3) G. Salton and M. J. McGill: Introduction to Modern Information Retrieval., McGraw-Hill, (1983)
- 4) ChaSen -2.4.2-- 形態素解析器
<http://chasen.naist.jp/hiki/ChaSen>
- 5) Yahoo!デベロッパーネットワーク - 検索 - ウェブ検索
<http://developer.yahoo.co.jp/webapi/search/websearch/v2/websearch.html>
- 6) Yahoo! Search blog: Our Blog is Growing Up ?
<http://www.ysearchblog.com/archives/000172.html>