F-028

コンテンツフィルタリングの自動化手法
An Automated Filtering Method for Web Contents

池田 匡邦 * 矢崎 俊志 * 阿部 公輝 *
Masakuni Ikeda Syunji Yazaki Kôki Abe

1. はじめに

コンテンツフィルタリングは，Web アクセスを監視し，コンテンツに応じてアクセスの可否を判断する技術である．しかし，既存の方法では，既知のコンテンツにしかフィルタリングを適用できないものが多い．そのため，コンテンツリストを更新し続けなければならず，多大なコストが必要になる．また，リストの更新を人手で行うことによって生じるフィルタリング精度の誤差も問題である．未知のコンテンツに適用できる方式もあるが，意味解析を用いるため，複数言語の混在する Web コンテンツには適さない．

既存のコンテンツフィルタリング方式として，ブラックリスト方式，ホワイトリスト方式，レイティング方式，キーワード方式がある．現在，ブラックリスト方式を用いた手法が一般的に用いられている．この方式では，問題のある Web サイトの URL を集めたリストを作成しておき，対象のサイトを無条件でブロックする．予め作成されたリストのみを対象とするため，フィルタリング時に誤判断が生じないメリットがあるが，未知の Web サイトには適用できない．また，フィルタリング漏れを防ぐためにはリストを頻繁に更新し続ける必要がある．

本研究では，これらの問題を解決するために，リンク構造解析を利用して，Web コンテンツリストの自動拡張と更新を行う手法を提案する．

2.Web サイトの分類手法

Web コンテンツリストの作成, 拡張を自動化するために，Web サイトを分類することは密接に関連する.Web サイトの分類手法には，Web コンテンツマイニングと Web 構造マイニングがある.Web コンテンツマイニングでは，Web サイトのテキスト情報を，自然言語処理等の技術を用いて分類する．対して，Web 構造マイニングでは，Web サイトのリンク構造を用いて分類を行う．ここでは，Web 構造マイニングの技術について示す．

HITS[1] は，リンク構造解析を用いた検索アルゴリズムである．サーチエンジンのキーワード検索などで与えられた Web ページ集合を，リンク構造に基づいて分類する．HITS では，Web 集合から 2 つの情報 (オーソリティとハブ) を抽出する．オーソリティは，トピックに関する情報量が多いことを示し，ハブは，オーソリティへのリンクが多いことを示す指標である．被リンク数が多いほどオーソリティの評価は高くなり，リンク数が多いほどハブの評価は高くなる．つまり，オーソリティの評価が高いサイトは優良なサイト，ハブの評価が高いサイトは優良なリンク集といえる．一般的な Web ページは，他のページをリンクし，又，他のページからリンクされているので，オーソリティとハブ両方の性質を持つ．

あるページ a に対するオーソリティとハブの値は，それぞれ以下の式で求めることができる.a → b は，ページ a からページ b へのリンクが存在することを示す．

$$authority(a) = \sum_{b, b \to a} hub(b)$$

$$hub(a) = \sum_{b, a \to b} authority(b)$$

3. 関連研究

Web ディレクトリは，トピック毎に Web サイト集合がリスト化されているので，コンテンツフィルタリングの為の Web コンテンツリストとして用いることができる.Web 構造マイニング技術を用いた Web ディレクトリの自動拡張手法 [3] では，Web サイトの参照関係を利用して，Web ディレクトリを自動拡張する．この手法ではキーワードによるロボット検索を用いて収集した関連 Web サイトをノードとする大域 Web グラフを作成する．このグラフに対して得られるオーソリティスコアを関連度とし，関連度の高い Web サイトを，Web ディレクトリに加える．

この手法では，キーワードを使って既存の Web ディレクトリの拡張を自動で広範囲に行うことを目的としている．本研究では，この手法を参考にしているが，キーワードを用いずに既存の Web ディレクトリをシードとして用いて同じトピックを持つ Web ディレクトリを自動的に拡張する．

4. 提案手法

Web コンテンツリストの収集に HITS を利用することで，リストの自動拡張と更新を行えるような手法を提案する．本手法はリンク構造解析を利用して，Web コンテンツリストの自動拡張と更新を行う点では，関連研究 [3] と同じであるが，関連 Web ページ検索アルゴリズムを用いて，既に存在するブラックリストのリンクから新たなブラックリストを生成することで，リストの自動拡張を行う．本手法のアルゴリズムは次の通りである．

1. 既知の Web ページをシードとして，関連 Web ページリストを作成する．
2. 関連 Web ページリストに対して HITS を適用し，オーソリティを求める．
3. 各関連 Web ページをシードとして，その関連 Web ページリストを求め，同様にオーソリティを求める．

*電気通信大学, The University of Electro-Communications

Copyright © 2011 by Information Processing Society of Japan and
The Instiute of Electronics, Information and Communication Engineers
All rights reserved.

4. 元の Web ページと各関連 Web ページのリストを比較し，共通したオーソリティを持つものを見つける．

5. 元のシードと発見した関連 Web ページリストのシードを合わせて，新たなシードとする．

これを繰り返すことによって，同一のトピックをもった Web ページを自動収集することができる．

提案手法のアルゴリズムの実行例を図 1〜4 に示す．図では，Step1，2 において，入力 Web ページをシードとして，関連 Web ページリストを求め，関連リストのオーソリティページを求めている．Step3 では，関連 Web ページの関連リスト①，②，③を求めている．Step4 では，関連 Web ページの関連リスト②のオーソリティが，元の関連リストのオーソリティと一致していることを示している．(関連 Web ページの関連リスト①，③のオーソリティは元の関連リストのオーソリティと一致しない．) Step5 では，関連リスト②のシードを，既に存在する入力データリストに加えている．

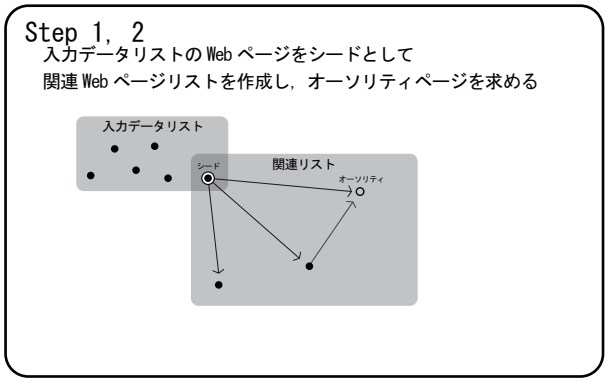

図 1: 提案手法のアルゴリズム実行例 Step1，2

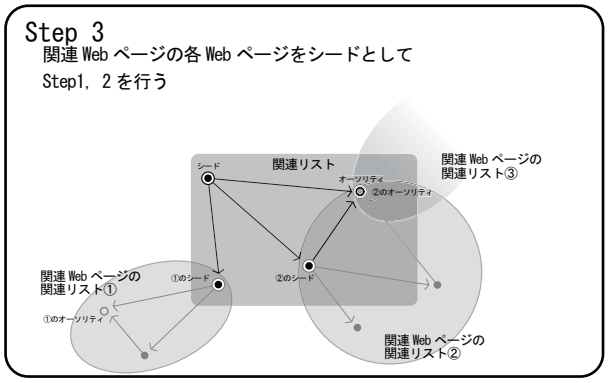

図 2: 提案手法のアルゴリズム実行例 Step3

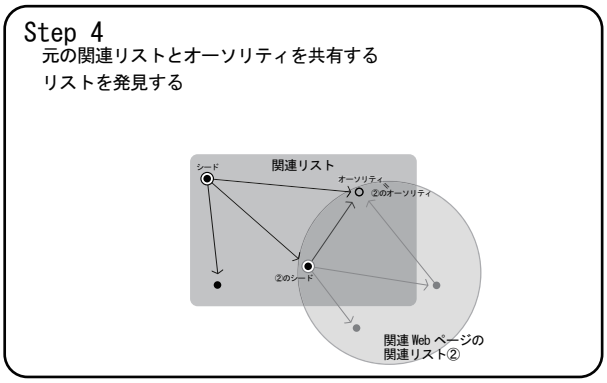

図 3: 提案手法のアルゴリズム実行例 Step4

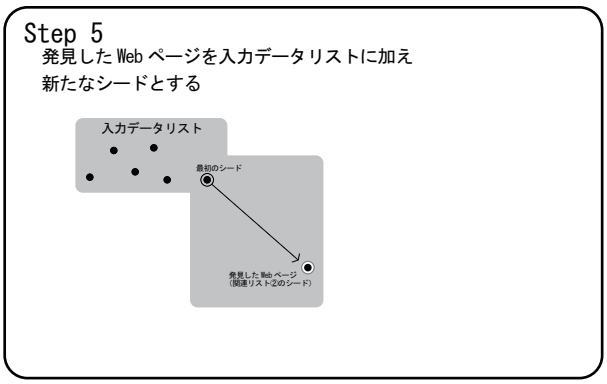

図 4: 提案手法のアルゴリズム実行例 Step5

5. 実験

提案手法を実装し，入力した Web ページから新たな Web ページを出力する実験を行った．実験では，既知の Web ページを複数個入力し，それぞれの関連 Web ページリストを作成して，同一のオーソリティを持つ Web ページを発見できるかどうかを確かめた．関連 Web ページリストを求める際，入力した HTML の $\langle a \rangle$ タグをリンクとして抽出した．既知の Web ページとして，表 1 に示す Google のディレクトリサービスに登録されているカテゴリリスト (Google ディレクトリ＞スポーツ＞サッカー＞ J リーグ) の 204 個の URL を用いた．表 1 には，各 URL に順番に ID を示してある．表は，入力に用いた 204 個の URL の内，はじめの 31 個とおわりの 16 個だけを示してある．http://や https://も省略してある．

6. 結果と考察

提案手法を適用して得られた結果を表 2 に示す．表 2 では，表 1 のシード URL を入力したとき，ID ごとに発見した URL の数と，発見した URL の例を示してある．発見できなかったものは除いてある．

たとえば，ID=26 の footballcs.info をシードとして提案手法を適用した結果，www.consadole.net/football 等が得られた．これは，footballcs.info の関連リストと，その中の www.consadole.net/football の関連リストがオーソーリティを共有していることを意味する．このよ

Copyright © 2011 by Information Processing Society of Japan and
The Instiute of Electronics, Information and Communication Engineers
All rights reserved.

表 1: 入力した SeedURL

ID	URL
0	www.albirex.co.jp
1	members.at.infoseek.co.jp/albij1up
2	www.ventforet.co.jp
3	www.sannichi.co.jp/VFK
4	vissel-kobe.co.jp
5	page.freett.com/starplutinum/photo-vissel-main.htm
6	e06126162.s44.xrea.com
7	kobe12.jp
8	kangaeru.fc2web.com
9	zeniya.hp.infoseek.co.jp
10	www.kataller.co.jp
11	www.geocities.co.jp/Athlete-Athene/8010
12	www.ssl.ykk.co.jp/ykkapfc/sunbbs
13	m-space.jp/a/?since1990alos
14	www.gamba-osaka.net
15	www.fcosaka.com
16	homepage3.nifty.com/aonoheya/sacgamba.htm
17	f44.aaa.livedoor.jp/ gamba/top.html
18	newwave-k.co.jp
19	wavyweb.lolipop.jp
20	www.consadole-sapporo.jp
21	www.phoenix-c.or.jp/ hiro/conindex.htm
22	www.consadeconsa.com
23	www.consadb.net
24	www.consadole12.com
25	www.geocities.jp/consaboraconsa
26	footballcs.info
27	www.alles.or.jp/ manami/consa
28	www.sagantosu.jp
29	www.synapse.ne.jp/ aisha/tosu
30	la-tosu.com
...	
188	www.tochigisc.jp
189	www.shimotsuke.co.jp/sports/t_sc
190	www2.ucatv.ne.jp/ nionio.sea/tochigisc
191	jfl-kikou.seesaa.net
192	sports.geocities.jp/tochisapo12
193	www.reysol.co.jp
194	www.shu.com/reysol
195	ouenniikou.at.infoseek.co.jp/index.htm
196	members.jcom.home.ne.jp/hellyeah/index.html
197	nagoya-grampus.jp
198	www.chunichi.co.jp/chuspo/article/grampus
199	gratube.shisyou.com
200	www.infobb.com/gran
201	plaza.harmonix.ne.jp/ k-suzuki
202	homepage2.nifty.com/n_grampus
203	homepage2.nifty.com/junkissa

うにして，204 個の URL をシードとして，新たに 1322 個の URL が得られた．また表 1 のシード URL を提案手法に入力していったとき，それまでに発見した関連 Web ページ数の累積数を図 5 に示す．

ある URL をシードとして，提案手法を適用して得られた結果が，入力と同じ URL になる場合がある (例えば ID=15 等)．また，シードとしてあるページを入力したとき同一サイト内のページが得られる場合がある (例えば ID=0 等)．また，本手法では，HTML の $\langle a \rangle$ タグをリンクとして抽出しているため，HTML 以外の言語で記述されたページでは，入力に対して共通のオーソリティを持つ Web ページが得られない．

実験から，本手法によって共通のオーソリティを持つ Web ページを得られることが示された．しかし，オーソリティの一致する関連 Web ページが，元のシード Web ページと同じトピックを持つかどうかは本手法では判断できない．この判断は，オーソリティの一致する関連 Web ページがシード URL 集合に属するかどうかを調べることで可能であると考えられる．Web コンテンツリストの自動拡張における本手法の有効性を示すには，このような実験をさまざまなシード URL 集合を用いて行うことが必要である．

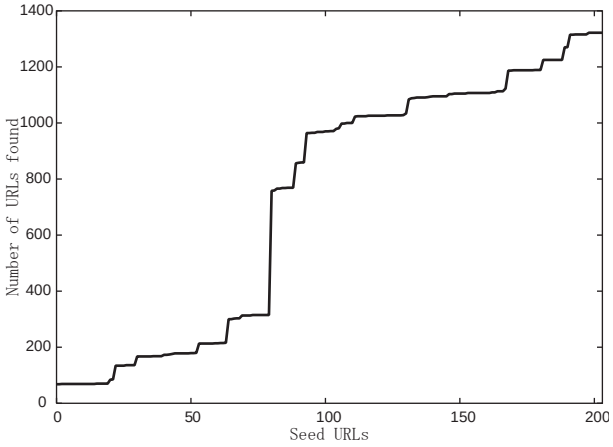

図 5: 発見した関連 Web ページの累積数. 横軸は表 1 のシード URL の ID を表す.

7. おわりに

Web コンテンツリストを HITS を利用して自動拡張と更新を行うための手法を提案した．実験により 204 個のシード Web ページからオーソリティの一致する関連 Web ページが合計 1322 個得られた．オーソリティの一致する関連 Web ページが，元のシード Web ページと同じトピックを持つかどうかの判断および，トピックと関係のないリンクの除去については今後の課題である．

参考文献

[1] J. M. Kleinberg : Authoritative Sources in a Hyperlinked Environment, Proc. ACM-SIAM Symposium on Discrete Algorithms, 1998, pp.668-677, 1998.

[2] J. Dean : Finding Related Pages in the World Wide Web, WWW8/Computer Networks, Vol.31(11-16), pp.1467-1479, 1999.

[3] 原田昌紀, 風間一洋, 佐藤進也 : 参照共起分析の Web ディレクトリへの適用, 情報処理学会情報学基礎研究会報告, Vol.2001-03-05, pp.45-52, 2001.

Copyright © 2011 by Information Processing Society of Japan and
The Instiute of Electronics, Information and Communication Engineers
All rights reserved.

表 2: オーソリティの一致した関連 Web ページリスト

ID	発見した URL の数	発見した URL の例
0	68	www.albirex.co.jp/info/sitemap.html
2	1	www.nisca.co.jp
15	1	www.fcosaka.com
20	14	www.consadole-sapporo.jp/funclub
21	1	www.s-pure.jp
22	49	www.consadeconsa.com/schedule
26	2	www.consadole.net/football
30	31	peroperon.seesaa.net/article/35235175.html
36	1	www.sanfrecce.co.jp/support/ticket/kazoku.html
40	5	jette.exblog.jp
42	1	ameblo.jp/pride-of-chiba
43	2	www.jubilo-iwata.co.jp
44	2	sankouchou.com/2010/08/869.html
50	1	www.geocities.co.jp/Athlete-Sparta/8090/hajime.htm
52	1	yaplog.jp/jubiliving
53	33	jubilation2.seesaa.net/article/115776754.html
59	1	www.jsgoal.jp
61	1	www.oms.co.jp
63	1	www.forza-fagi.com
64	84	www.vegalta.co.jp/index.html
66	2	www.cgiboy.com
67	1	www.pressart.co.jp/s-style
69	10	vegalta.rulez.jp
73	2	www.kahoku.co.jp/spe/spe106/index.htm
80	443	www.montedio.or.jp/info
81	1	plaza.rakuten.co.jp/pekomikuchi
82	7	www.montedio.org/modules/contact
84	2	www.jsgoal.jp/special/2010j2book
86	1	d.hatena.ne.jp/rosso-penya
89	87	www.ehimefc.com/fanclub_login
90	2	blog.livedoor.jp/arancino_kanto
91	1	www.geocities.jp/ehimesta
93	105	www.urawa-reds.tv/rnn
95	1	www.komura.com/urawa
97	3	uragi.com/page2.html
100	2	www.tajimans.com/godzilla
102	1	urawamaniac.com
104	8	redsoul.net/archives/cat_113.html
105	2	rj2001.net/we_are_reds
106	17	red-quest.seesaa.net/index-2.html
108	2	www.geocities.jp/welcome_to_redhell/top2.html
111	23	www.f-marinos.com/index.html
112	1	www.ardija.co.jp
116	2	www.nasudaisuke.com
123	1	www.soccer-kentei.jp/blog/top
129	1	www.sanga-fc.jp/others/inquiry/form
130	6	www.sanga-saporen.net/modules/news
131	50	marostyle.seesaa.net/article/211151516.html
132	4	megawave783.blog96.fc2.com
133	1	www.kotsu-kotsu.jp
134	2	www.so-net.ne.jp/antlers/fanzone/catalog/index.html
138	1	ekitan.com
139	2	www12.ocn.ne.jp/ kyoshika
140	1	blog.livedoor.jp/moichi9377/archives/52301133.html
146	8	www.bellmare.co.jp/tickets
148	2	d.hatena.ne.jp/neobellmare
153	2	www.fcmito-crazy.org
162	2	www.frontale.co.jp/index.html
164	4	frontale.net/modules/d3pipes
167	10	uhauha.jp/archives/2010/03/3-27.html
168	64	tdrk.blog4.fc2.com/blog-date-200906.html
170	1	www.movi.co.jp
178	1	www.vasagey.com
181	36	www.verdy.co.jp/sitemap-2
189	45	www.hokkaido-np.co.jp
191	45	jfl-kikou.seesaa.net/article/127343278.html
193	1	blog.reysol.co.jp/news/topteam
198	6	cgi2.chunichi.co.jp/tko/tochu/modama/modama_form.shtml

Copyright © 2011 by Information Processing Society of Japan and
The Instiute of Electronics, Information and Communication Engineers
All rights reserved.