1G2-01

コンテンツ情報を用いたWebコミュニティの洗練

Refinement of Web Communities using Web content information

石原 達生　松井 藤五郎　大和田 勇人
Tatsuo Ishihara　Tohgoroh Matsui　Hayato Ohwada

東京理科大学 理工学研究科 経営工学専攻
Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science.

The rapid growth of the World Wide Web has created many challenges for Web mining, such as general purpose crawling, search engines, web directories, classifing web pages, and discovering Web communities. Web structure mining is one of the most major fields of Web mining. We focus on discovering Web communities. Murata has built a system which discovers complete bipartitie graphs based on a assumption that hyperlinks to related Web pages often co-occur. Although this system succeeds in discovering several Web communities, it causes a serious problem called 'topic drift problem' frequently. This paper proposes a method for refining of Web communities using the content. Experimental results show that our method is effective for resolving topic drift problem.

1. はじめに

近年，Web 上のデータは爆発的に増加し続けており，その中からユーザが必要な情報を獲得することは非常に困難になってきている．現在，ユーザが Web から必要な情報を獲得する際にはサーチエンジンが用いられる．サーチエンジンは，全文検索型とディレクトリ型の 2 種類に大きく分類できる．全文検索型サーチエンジンによる検索は，ロボットが自動的に収集したページを対象にキーワードを入力することにより検索を行うが，自然言語には数多くの多義語や同義語があるために，ユーザが入力したキーワードだけから情報を検索することは困難であり，ユーザが必要とする情報を提供することができない．ディレクトリ型サーチエンジンは，関連する内容のサイトを収集することは可能だが，人手で分類するために膨大な量のサイトをカテゴライズすることは非常に困難な作業である．また，複数のトピックにまたがっているサイトも少なくないため，あるサイトと同じ興味を持ったサイトだけを検索することが難しい．

以上のような理由から，ハイパーリンクのグラフ構造に基づく Web structure mining において，興味を共有する Web ページ集合 (Web コミュニティ) を発見するための研究が行われてきている．村田は，参照の共起性に基づく Web コミュニティを発見する手法を提案した [村田 01]．しかし，村田の手法には，発見される Web コミュニティのトピックがずれてしまうトピック・ドリフト現象と呼ばれる深刻な問題がある．

本論文では，村田の手法のような完全 2 部グラフ構造を基本的な構造としてもつ手法に起こりやすいトピック・ドリフト現象に対し，Web ページのコンテンツ情報を積極的に利用することで抑制する手法について述べる．

2. 村田の Web コミュニティの発見手法

村田の手法は，ユーザから与えられた Web ページの URL 数個を基に，その URL を含んでいるような完全 2 部グラフを見出すことを目標としている．具体的には以下の処理を繰り返すことによって Web コミュニティの発見を行う．なお，以後の説明において，完全 2 部グラフ $K_{i,j}$ におけるリンク元の i 個の URL を *fans*，リンク先の j 個の URL を *centers* と呼ぶこととする．

連絡先: 石原達生，東京理科大学 理工学研究科 経営工学専攻 大和田研究室, 千葉県野田市山崎 2641, 04(7124)1501, j7404606@ed.noda.tus.ac.jp

2.1 *centers* を参照する *fans* の検索

入力された URL を *centers* として含んでいるような完全 2 部グラフを発見するために，入力 URL の全てに対してリンクを張っている Web ページをサーチエンジンの持つ機能である backlink 検索により獲得する．獲得した URL を *fans* とする．

2.2 *centers* の新たな URL の追加

得られた *fans* の URL に順次アクセスして HTML ファイルを取得し，各々のファイルに含まれているハイパーリンクの URL を全て抽出する．その中で最も出現回数の多いものを *centers* に追加し，新たな *centers* について上述の処理を繰り返す．一般に AND 検索の条件が増えると獲得される URL の個数は徐々に減少していく．*fans* の個数があらかじめ定めた個数になったら終了する．

3. 村田の手法における問題点

3.1 トピック・ドリフト現象

2 部グラフ構造を基本的な構造として持ち，Web ページ間の重要性や関連性を解析して利用するような手法は，目的するページと関連性のないと思われるページが関連するページとして選択されるという現象が現れることがある．このような現象をトピック・ドリフト現象と呼び，以下のような原因で発生する．

1. リンクの意図を考慮しないで，一括してハイパーリンク情報を処理すると，一般的なトピックや著明サイトほど高く評価されやすくなる．
2. Web 文書のリンク構造は複雑であり，プログラムが機械的に生成した大量のリンクの影響をうけることがある．

[村田 01] の手法においても 2 部グラフ構造を持つため，このトピック・ドリフト現象が発生しやすいという問題がある．

3.2 初期の入力

[村田 01] は初期の入力には必ず 2 つ以上の組合せで入力する必要がある．その時，初期の入力 URL に，実際はあまり関連性の無い URL を組合せてしまうと，多くの場合 1 つも *fans*

を獲得できず，処理がすぐに終了してしまう．初期 URL を 1 つで入力すると高い確率でトピック・ドリフト現象が発生し，全く関係の無いページが含まれていることが本研究の予備実験により確認された．ユーザがある初見のページを入力したい時にそのページに関連する他のページを捜し出すのは困難である．そこで，1 つの URL からトピックずれの少ない Web コミュニティを発見することが必要となる．

4. Web コミュニティの洗練手法

本節では，上述の問題点を踏まえ，本手法における Web コミュニティの洗練手法について記述する．

本システムの一連の流れを示す (図 1)．まず初めにユーザが Seed URL を入力する．その Seed URL のコンテンツ情報を取得して，TF・IDF 値を計算し (1)(2)(3)，後に類似度を計算するために保存しておく (Step1)．

- 相対索引語頻度 (relative term frequency)

  $tf(t)$: ある文書 d 中に出現する索引語 t の数.

  w_t^d: 文書 d 中に出現する索引語 t の相対頻度

$$w_t^d = \frac{tf(t,d)}{\sum_{s \in d} tf(s,d)} \tag{1}$$

- IDF (Inverse Document Frequency)

  N: 全文書数,

  $df(t)$: 索引語 t が出現する文書数.

$$idf(t) = \log \frac{N}{df(t)} + 1 \tag{2}$$

- TF · IDF

$$tf \cdot idf(t) = w_t^d \cdot idf(t) \tag{3}$$

ここでの入力はユーザの利便性を考えて，一つの URL とする．次にこの URL よりサーチエンジン (今回は AltaVista) を使用して，backlink 検索を行う (Step2)．backlink 検索とは今回の例で説明すると，Seed URL に対してリンクを張っているようなページを検索するということである．その検索結果から上位 N 件を取得 (本実験では 50 件) をして，それを *fans* とする (Step3)．50 件を取得できない時は処理を終了する．

次に，各 *fans* のファイルにアクセスしてハイパーリンクを取得し (Step4)，その抽出されたリンクの出現回数の多い順にソートする．そして，上位 M 件 (本実験では 20 件) を取得する (Step5)．そのソートされたリンクのコンテンツ情報を取得して，TF・IDF 値を計算する (Step6)．

次に，その抽出された 20 件のページと元の *centers* との類似度を計算する (Step7)．

- 余弦

  TF・IDF で数量化された文書のベクトル x_i, y_i の類似度.

$$\sigma(d_x, d_y) = \frac{\sum_{i=1}^{T} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{T} x_i^2 \times \sum_{i=1}^{T} y_i^2}} \tag{4}$$

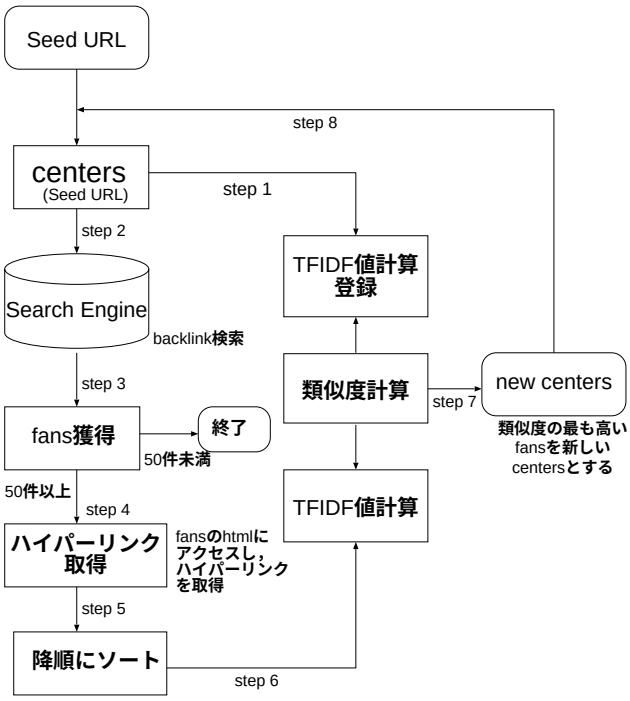

図 1: システム構成図

最も高い類似度をもつページを新しく *centers* に追加する (Step8)．ただし，Step7 の 2 回目のループにおいては，*centers* の数が 2 個存在するので，2×20 回の類似度の計算をすることになり，20 件獲得した *centers* 候補については各候補が 2 つの類似度をもつことになる．よって，その場合はその 2 つの *centers* との類似度の積をその *centers* 候補の類似度とする．以上のようにして，*centers* との類似度が最も高い候補を新しく *centers* に加えるものとする．

以上の処理を繰り返すことにより Web コミュニティの洗練を行う．

5. 実験

本システムの有効性を確かめるために，Java を用いてシステムを構築し実験を行った[*1]．Seed URL は，日本経済新聞の大学生の就職希望ランキングに掲載されている業種別ランキングから，任意の業種を選択しその中から任意に抽出した企業の URL とした．ここでは，旅行・レジャー，輸送用機器，官公庁・農協・その他団体，銀行，精密機械・その他製造，電子・電機の 6 業種から URL を選択した．

表 1 に，旅行・レジャーの分野の Seed URL を入力した時に獲得された新しい *centers* を示す．左の番号は，*centers* が追加された順番を表す．

表 2 は，各ジャンルに対して，本手法と村田の手法のトピックずれの有無を比較したものである．

*1 比較に用いた村田の手法は [村田 01] に基づいて，本システムと同様に実装した．このシステムは [村田 03] と同様の結果が得られることを確認している．

表 1: URL の出力の例 (旅行・レジャー)

www.jtb.co.jp の場合

	村田の手法	本手法
1	www.jtb.co.jp	www.jtb.co.jp
2	www.japan-guide.com	www.japan-guide.com
3	www.japantimes.co.jp	www.japantimes.co.jp
4	www.jnto.go.jp	www.jnto.go.jp
総数	4	4
トピックずれなし数	4	4

www.his-j.com の場合

	村田の手法	本手法
1	www.his-j.com	www.his-j.com
2	www.arukikata.co.jp	www.arukikata.co.jp
3	www.nikkei.co.jp	www.jtb.co.jp
4	www.asahi.com	www.nta.co.jp
5	www.yomiuri.co.jp	
6	www.mainichi.co.jp	
7	www.mapion.co.jp	
総数	7	4
トピックずれなし数	2	4

www.nta.co.jp の場合

	村田の手法	本手法
1	www.nta.co.jp	www.nta.co.jp
2	www.google.co.jp	www.tour.tokyu.com
3	www.ascii.co.jp/ghelp	www.knt.co.jp
4	www.jnto.go.jp	www.jtb.co.jp
5		www.ana.co.jp
6		www.jal.co.jp
7		www.his-j.com
8		www.mytrip.net
9		travel.yahoo.co.jp
総数	3	9
トピックずれなし数	1	9

6. 考察

JTB (www.jtb.co.jp) のページを入力した時は両者は全く同じ出力結果を得た．これは，村田の手法においてもトピックずれが認められないために，生じた結果であると考えられる．

HIS (www.his-j.com) のページを入力した時は，村田の手法の方が多くのページが出力された．しかしながら，途中で日本経済新聞社 (www.nikkei.co.jp) のページが追加されてしまったことにより，それ以後は，ニュース系のページになってしまっている．一方，本手法の出力結果を見ると，ページの数は少ないものの，トピックずれが防がれていることが分かる．

日本旅行 (www.nta.co.jp) のページ を入力した時も，出力結果をみると，村田の手法は明らかにトピックがずれており，そのために出力結果も極端に少なくなってしまった．これは，初めに追加された URL が Google (www.google.co.jp) のページであり，トピックが極端に違っていたことが原因と考えられる．これに対し，本手法は，多くの旅行系のページを網羅している．

また，表 2 で本手法と村田の手法を比較すると，村田の手法が 6 ジャンル中 3 つずれているにもかかわらず，本手法は 6 ジャンル中 1 つしかトピックがずれを起こしていない．以上のことから本手法の方が明らかにトピックずれを防いでいると言える．

また，本手法を適用して Web コミュニティの洗練を行うと，Web サイトの獲得総数が多くなるということが分かった．本手法と，村田の手法を比較すると全てのジャンルにおいて，本手法のほうが獲得総数が多いことが分かる．このことにより，本手法を用いてよりトピックずれのないページを順番に追加していくことで，より多くのページを取得できることが明らかになった．より多くのページを獲得することはユーザにとっても知識獲得の際に非常に役に立つと考えられる．

7. 結論

本論文では，Web ページのコンテンツ情報を用いた手法で，Web コミュニティを洗練させる手法を示した．この手法を用いることで，1 つの URL の入力でも， [村田 01] と比較してトピックずれの少ない Web コミュニティに洗練することに成功した．また，本手法を適用し，Web コミュニティを洗練させていくことで，結果的により多くのページを獲得できることが実験より分かった．

今後は各 Web ページに対する効果的な重みづけの方法を考えたり，Web コミュニティの洗練過程でシステムとユーザが対話的に通信を行うようなシステムを構築することによりユーザの嗜好に合致するような Web コミュニティに洗練することが可能になると考えられる．

表 2: 各ジャンルの出力結果

村田の手法による各ジャンルの出力結果

ジャンル	トピックずれなし	獲得総数
旅行・レジャー	7	14
輸送用機器	5	7
官公庁・農協・その他団体	10	10
銀行	13	13
精密機械・その他製造	14	14
電子・電気	4	10

本手法による各ジャンルの出力結果

ジャンル	トピックずれなし	獲得総数
ホテル・旅行・レジャー	17	17
輸送用機器	6	8
官公庁・農協・その他団体	11	11
銀行	14	14
精密機械・その他製造	15	15
電子・電気	19	19

参考文献

[村田 01] 村田剛志, 参照の共起性に基づく Web コミュニティの発見, 人工知能学会論文誌 vol.16 no.3, pp.316-323, 2001.

[村田 03] 村田剛志, Web コミュニティの中心性, The 17th Annual Conference of the Japanese Society for Artificial Intelligence, 2003.

[徳永 99] 徳永健伸, 情報検索と言語処理. 東京大学出版会, 1999.

[茶筅 2.3.3] 松本裕治, 北内啓, 山下達雄, 平野義経, 松田寛, 高岡一馬, 浅原正幸. 形態素解析システム 茶筅 version2.3.3 使用説明書,
奈良先端科学技術大学院大学.