テキスト情報を考慮した企業コミュニティの獲得

Acquiring community of companies using text information

石原 達生 † 松井 藤五郎 †† 大和田 勇人 ††
Tatsuo ISHIHARA† Tohgoroh MATSUI†† Hayato OHWADA ††

† 東京理科大学大学院 理工学研究科 経営工学専攻
Department of Industrial Administration, Graduate School of Science and Technology, Tokyo University of Science

†† 東京理科大学 理工学部 経営工学科
Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science

j7404606@ed.noda.tus.ac.jp {matsui, ohwada}@ia.noda.tus.ac.jp

従来，ある分野の企業を知るためには，業界地図などの書籍などによって調べるが，たくさんの企業が存在する現在では，Web 上からそのような情報を調べることが一般的になっている．本論文では，ある特定の分野の企業の集合である企業コミュニティを Web から獲得することを目的とする．類似した構造として Web コミュニティというものがあるが，Web コミュニティが一般的に興味を共有するページ集合であるのに対し，企業コミュニティとは，興味を共有する企業サイトの集合である．本論文では，そのような企業コミュニティをそれらのサイトに存在する Web ページのテキスト情報を有効に利用し，獲得する方法について述べる．

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．そのような背景から，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．Web コミュニティを発見する過程で生じてしまう問題にトピックドリフト問題というものが考えられる．これは，Web ページには複数の内容が記述されており複数の興味へのリンクが張られていることに起因する．

我々は，これまでにそのようなトピックドリフト問題を抑制し Web コミュニティを獲得する研究を行ってきた [1]．しかし，それらの獲得されたコミュニティは，単に Web ページの集合であり，あるジャンルの企業のコミュニティではなく，ある企業のトップページのコミュニティというものであった．

そこで本論文では，現実世界のある特定の分野の企業群を企業コミュニティとし，それを獲得することを目的とする．[1] のコミュニティ獲得ステップを基本的な獲得ステップとするが，今回はテキストの使用方法を改良し企業コミュニティが獲得できるかどうかの実験を行う．

具体的に，企業の’サイト’間の類似性を判定し，新しいサイトを企業コミュニティに追加するというステップをとるが，これは，企業のトップページのテキスト情報を用いるだけではなく，企業の自サイトへのリンクを取得し，使用することでこれを達成する．

2 Web コミュニティの獲得

本節では，まず企業コミュニティの獲得の基本的な方法となる [1] の手法について説明し，今回の提案する企業コミュニティの獲得方法との相異点を述べる．

2.1 Web コミュニティ獲得方法

[1] は，ユーザが 1 つ入力 URL を与え，そのページのテキスト情報を使用して，よりトピックずれの少ない Web コミュニティを獲得することを目的としている．これは，[2] と比較し，よりトピックずれの少ないコミュニティを獲得することに成功した．具体的には以下の処理を繰り返すことによって Web コミュニティを獲得していく．なお，このコミュニティを獲得する際に完全 2 部グラフ構造を用いるが，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．

2.2 テキスト情報

後に説明するが，[1] では *centers* に新しい要素を追加する際に類似度計算を行うが，その際に，テキスト情報を用いる．このテキスト情報とは，html ファイルからのプレーンテキストを形態素解析 [3] にか

け，単名詞のみを抽出し，それを索引語としたものである．索引語は，TFIDF によって重み付けを行う．

2.3 *centers* を参照する *fans* の検索とテキスト情報の獲得

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．また，入力した URL のテキスト情報を抽出し，後で *centers* に登録する際に類似度を計算するために登録しておく．

2.4 類似度計算による *centers* への新たな 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から上位 N 件を獲得し，*centers* との類似度を計算する．その中で類似度が最大のものを *centers* に追加し，新たな *centers* について上述の処理を繰り返す．この獲得ステップは，*fans* の個数があらかじめ定めた個数になったら終了する．

2.5 問題点

上述の方法は，多くのトピックずれの少ない Web コミュニティを獲得することに成功し，backlink 検索を行うことで，通常，同業他社などからは直接リンクは張られていないにもかかわらずそのような企業の URL を獲得できるというメリットを持っているが，獲得できたコミュニティは，単純に企業のトップページのコミュニティであり，本質的には，現実世界の企業のコミュニティとは言えるものではなかった．

また，使用するテキスト情報であるが，これは，1 つ企業に対して 1 つの html ファイル (トップページ) を用いていたために，テキストの情報量が非常に少ない事例もあった．これは，近年トップページにはフラッシュなどの動的なコンテンツが多く含まれることに起因すると考えられる．よって，現実世界のコミュニティである企業コミュニティを獲得するためには，企業のトップページの html ファイルのみを使用するだけではなく，ある程度，その自サイト内の他のテキスト情報もより有効に使用することが必要であると考えられる．

次節では，上述の問題点を踏まえて企業コミュニティの獲得方法について説明する．このステップは [1] とほぼ同様のものであるが，細部が多少異なっているために，詳細な獲得ステップについても説明する．

3 企業コミュニティの獲得

Web コミュニティとは，一般的に興味を共有するページ集合であることに対し，本論文で述べる企業コミュニティとは，興味を共有するサイトの集合である．例えば，コンピュータの Web コミュニティとは，コンピュータに関する内容が記述されているような Web ページの集合である．それに対し，コンピュータの企業コミュニティとは，NEC や日立などといったような，コンピュータメーカのサイトの集合である．

本節では，テキスト情報を利用し，企業コミュニティを獲得する方法を具体的に述べる．まずは，使用するテキスト情報についての説明をし，次に具体的に獲得ステップを記述する．

3.1 テキスト情報

企業コミュニティの獲得の際に使用するテキスト情報は，企業のサイトに存在する html ファイルのプレーンテキストである．そのプレーンテキストを形態素解析器にかけ，単名詞を抽出する．また，その際に複合名詞も抽出し，それぞれをそのテキストの索引語とする．索引語の重み付けは，TFIDF を用いた．形態素解析器としては，茶筌 [3] を使用し，複合名詞の抽出としては東京大学中川研究室・横浜国立大学森研究室で開発された用語抽出システム [4] を使用した．今回は，中川らのシステムにより付与される用語のスコアについては考慮せず，単純に複合名詞を抽出するために使用した．

具体的な獲得ステップに関しては，次で詳しく説明するが，上述のテキスト情報は，企業コミュニティに新しい要素を追加するステップにおいて類似度を比較する際に使用される．尺度としては，コサイン類似度を用いて類似性の判定を行った．

3.2 獲得ステップ

本手法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 と自サイトへのリンクを取得し，テキスト情報を抽出する．そして，それらの TFIDF 値を計算する．*centers* の

自サイトへのリンクはいくつか存在するが，類似度計算の単純化のために，最終的には全てのリンクを 1 つのファイルとして扱うことにし，後に類似度を計算するために保存しておく (Step1).

Seed URL は 1 つを与え，サーチエンジン (今回は AltaVista) を使用して，backlink 検索を行う (Step2). その検索結果から上位 N (本実験では 50 件) 件を取得をする (Step3). *fans* を N 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 (1 次候補) を取得する (Step5)．本実験では，$M = 30$ で実験を行った.

次に *centers* の時と同様に 1 次候補から自サイトへのリンクを全て取得し，で TFIDF 値を計算する (Step6)．ここでも，*centers* の時と同様，類似度計算の単純化のために，1 サイトにつき 1 つのファイルとして扱うことにした.

次に，一次候補と元の *centers* との類似度を計算する (Step7).

次に，閾値 R 以上 (本実験では $R = 0.7$) の類似度をもつページ (2 次候補) 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，$2\times M$ 回の類似度の計算をすることになり，1 次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* との類似度の和をその *centers* 候補の類似度とする．よって，*centers* と 1 次候補の類似度 $\sigma(Ce, Ca)$ は以下で表すことができる.

$$\sigma(Ce, Ca) = \sum_{i=1}^{C} \frac{\sum_{j=1}^{T} ce_j \cdot ca_j}{\sqrt{\sum_{j=1}^{T} ce_j^2 \times \sum_{j=1}^{T} ca_j^2}} \qquad (1)$$

ここで，ce_j, ca_j はそれぞれ，centers, 1 次候補の索引語 j に TFIDF で重み付けされた値であり，T は索引語の総数，C は *centers* のサイト数である.

以上のようにして，2 次候補から新しく *centers* に加えるものとする．2 次候補が複数個存在したときは (大抵始めは存在する)，SeedURL と 2 次候補のペアでコミュニティの獲得を行う．獲得の途中で重複するページが別のルートにも存在した場合は，獲得できたコミュニティ同士を接続する.

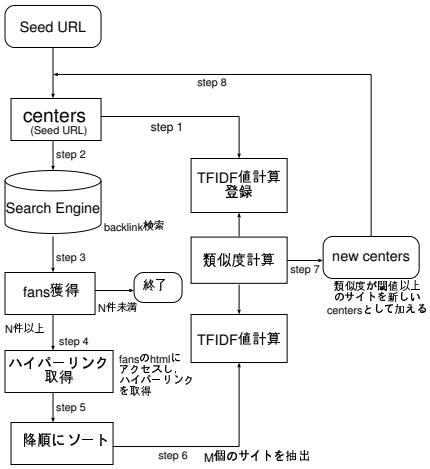

図 1: 処理の流れ

以上のような greedy なアプローチで処理を繰り返すことにより企業コミュニティの獲得を行う.

4 実験

企業コミュニティの獲得ができるかどうかを確かめるために実験を行った．Seed URL は日立 (http://www.hitachi.co.jp/) のページとする．獲得できた企業コミュニティの結果を表 1 に示す.

表 1: 企業コミュニティ出力の例

企業コミュニティ
http://www.nec.co.jp/
http://www.sharp.co.jp/
http://www.nissan.co.jp/
http://www.sony.co.jp/
http://www.kyoto-u.ac.jp/
http://www.bmw.com/

5 考察

この結果をみると少ないながらもある程度はコンピュータメーカの同業他社サイトをいくつか獲得することができている．なお，順位は特に意味を持たない．しかし，サイトの中にいくつか別分野の企業が追加されてしまっている．これらの企業は獲得のステップにおいて，最後の方にやむをえず追加されてしまったためであると考えられる．実際，これらの類似度を見るとそんなに高くない値であった．あま

り，多くの企業コミュニティが獲得できなかった理由として，1 次候補の選択方法に問題があると考えられる．実際，一番最初の (http://www.hitachi.co.jp/) を入力した際の 1 次候補のリストをトレースすると，ほとんど (26/30) が海外のサイトであり，この時点で既に同じジャンルに属する企業コミュニティとは言いがたい．よって，この 1 次候補の選択方法についてより考える必要がある．

また，本実験では主に日立の同業他社のコミュニティを獲得するための実験であったが，日立のグループ企業についても獲得できるかの考察を行う．日立のグループ企業に関しては，同業他社の企業と異なり実際に直接リンクをしている可能性が高いために，http://www.hitach.co.jp/の URL を 1 つ入力した際の *fans* のリストを取得し，結果を表 2 に示す．なお，この結果は，ドメインに'hitachi' という文字が入るものだけを抽出し，その総数は *fans* の総数 $N(= 50)$ 個中 19 個を獲得した．この結果だけでも，同グループのサイトは単純に fans を取得することで多くのコミュニティを獲得可能であることがわかる．

表 2: http://www.hitachi.co.jp/を入力した際の *fans* のリスト

fans のリスト
http://direct.hitachi.co.jp
http://www.hitachijoho.com
http://www.hitachi-hitec.com
http://www.hitachidisplays.com
http://www.hitachi-hic.jp
http://floracity.hitachi.co.jp/go/prius
http://www.hitachiacs.co.jp
http://www.hitachi-kenki.co.jp
http://www.hitachiplant.hbi.ne.jp
http://www.hitachi-densa.co.jp
http://www.hitachi-hb.co.jp
http://www.hitachi-it.co.jp
http://www.hitachi-hec.co.jp
http://www.hitachi.ca
http://www.hitachimetals.com
http://www.hitachigst.com/portal/site/jp
http://www.hitachi-kizai.co.jp
http://www.hitachi-ies.co.jp
http://www.hitachi.co.jp/Prod/vims/mobilephone
総数：19 個

6 関連研究

豊田らは，定期的に収集した大規模なデータから主要なコミュニティを抽出し，それらの相関図を構築し，アーカイブ間でコミュニティの比較を行うことで時系列的変化を抽出する手法を提案した [5]．このコミュニティの相関図はコミュニティチャートと呼ばれる有向グラフで表現され，多くのコミュニティを獲得することができている．彼らの手法は，部分的に見ると，多くの企業とそれに関連するページを獲得することができているが，被リンクの多いページを Seed Set として始めているために，ある特定のジャンルのみの検索ができない．実際に獲得されたコミュニティを見てみると，コンピュータ関連のコミュニティの回りには yahoo や vector などのサイトも確認することができる．Web 上の大規模な関連性を見る時はこのような情報は必要かもしれないが，我々は抽出精度の再現率という点よりも適合率という点に焦点を当てている．

7 結論

本論文では，各サイトに存在するテキストの情報を有効に利用することで企業コミュニティの獲得を行った．実験より，多くの企業コミュニティは獲得することは困難であったが，1 次候補中の少ない同業他社のサイトを獲得することに貢献したと言える．これは，テキストの情報量を増やしたことでより類似性の高いものを公正に判定することができたためであると考えられる．

今後は，考察で示したとおり，ある特定の企業の同業他社とそのグループ企業を取得することが十分できるという可能性が示されたので，Web 上のデータから業界地図を作成するようなシステムを構築する予定である．また，コミュニティが内包する意味を特定するためには，コミュニティ内部で成り立つような言葉を獲得するアプローチが有効であると思われるため，その問題に対しても取り組むことを検討している．

参考文献

[1] 石原達生, 松井藤五郎, 大和田勇人: コンテンツ情報を用いた Web コミュニティの洗練, 2004 年度人工知能学会全国大会 (第 18 回), 1G2-01 (2004).

[2] 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.

[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.

[4] 中川裕志、森辰則、湯本紘彰: 出現頻度と連接頻度に基づく専門用語抽出, 自然言語処理, Vol.10 No.1, pp. 27-45, (2003).

[5] Toyoda, M. and Kitsuregawa, M.: Creating a Web Community Chart for Navigating Related Communities, Proceedings of the 12th ACM Conference on Hypertext and Hypermedia (Hypertext 2001), pp. 103-112 (2001).