Wikipediaを類義語辞書として用いた企業クラスタリング

Company Clustering using Wikipedia as Synonym Dictionary

山田 裕文[*1] 松井 藤五郎[*2] 大和田 勇人[*2]
Hirofumi Yamada Tohgoroh Matsui Hayato Ohwada

[*1]東京理科大学 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

[*2]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

In this paper, we proposes a novel method of a company clustering. A past research cluster a company list by using text information at each relation to the company that inputs it first. However, to cluster by using text information in a past research, the inconsistent spelling has had a big influence on accuracy. We propose the method that uses the synonym dictionary that uses Wikipedia to obtain a company list. Experimental results show our method solve the problem of the inconsistent spelling.

1. はじめに

ある分野の企業を知るためには就職四季報などに存在するカテゴリサーチを用いて行う方法と，キーワードによって検索する方法がある．前者の場合，一つの企業について一つのカテゴリのみしか確認できないとう問題がある．近年，複数の事業を持つ企業も少なくないため，このような検索方法では難しい．後者の場合も多くは主要な事業のものしか記述されておらず，もし仮にすべての事業内容を記述すると膨大な量になるため現実的には不可能である．また，両者は共にある程度の予備知識を必要とすることから，これらの方法は現実的ではない．

大和田研究室では，テキスト情報を利用し，一つの企業を入力することで，その企業に関係のある企業のリストを入手し，リスト内の企業を入力企業の関係ごとにクラスタリングするという手法を提案した [takada 06]．これらは，各クラスタは入力企業との関係を示したタグが付与されていて，出力結果は視覚的に見やすくなっている．

しかし，テキスト情報を用いてクラスタリングを行うため，企業名がどのように表記されているかということが精度に大きな影響を与えてきた．精度が低下する原因は主に二つある．一つ目は URL から企業名に変換する際に，WHOIS サービスというドメインから登録してある企業名を求めるサービスを利用しているのだが，その登録されている企業名が正式でないこと．二つ目はテキストの内容によっては企業名に正式名称でなく通称が使われることである．実際には WHOIS で取得した企業名が正式名称か通称のどちらであるか分からないので，どちらかということによらず，最終的に正式名称と通称の両方を取得できることが望ましい．

そこで本研究では，大和田研究室での従来手法を拡張する．まず一つの企業を入力するとその企業に関係のある企業リストを入手し，その企業リストの企業に対して Wikipedia を用いた類義語辞書から企業名の正式名称・通称を取得する．そして，その正式名称・通称を含んだ新しい企業リスト内の企業を，入力企業の関係ごとにクラスタリングする手法を提案する．

連絡先: 山田裕文，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，j7409630@ed.noda.tus.ac.jp

2. Wikipedia を用いた類義語辞書

2.1 Wikipedia とは

Wikipedia は,Wiki をベースにした大規模 Web 百科事典であり，幅広い分野の記事（概念）を網羅している Wikipedia は，この幅広いトピックの網羅性以外にも，密なリンク構造などの興味深い特徴をいくつか持つ．また，URL により語彙の意味が一意に特定されている点や，リンクテキストの質の高さなども Wikipedia の特徴である．

2.2 類義語辞書

これら Wikipedia の持つ特徴は，知識抽出のコーパスとして極めて有利に働くことが各種の研究によりここ数年で急速に解明されてきた．中山らは，Wikipedia のページ間リンクの構造を解析することで，大規模で精度の高い類義語辞書を構築し，その結果を公開してきた [nakayama 07a, nakayama 07b]．これらは，ページ間のリンク構造を解析するものであるので，Wikipedia に登録されている記事（企業）なら, 企業名の正式名称・通称も取得することが可能である．よって，本研究では,Wikipedia を用いた類義語辞書を企業の正式名称・通称を取得することに使用する．

3. 提案手法

本研究の提案手法の全体の流れを図 1 に示す．

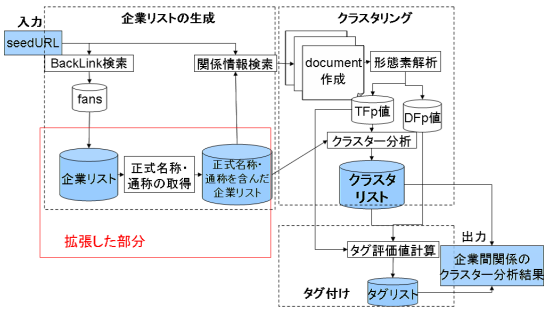

図 1: システム全体の流れ

3.1 正式名称・通称を含んだ企業リストの生成

企業リストの生成は従来手法と同様にまず企業リストを獲得する．そして，得られた企業リスト内の企業一つ一つに対して，Wikipedia を用いた類義語辞書*1から正式名称・通称を取得する．類義語辞書に入力するクエリを企業リスト内の企業名とし，類義語が取得できたならば，それを正式名称・通称とし，入力したクエリと得られた正式名称・通称を正式名称・通称を含む企業リストに加える．類義語が取得できなかった場合には，入力したクエリのみを正式名称・通称を含む企業リストに加える．

3.2 検索エンジンからの関係情報の取得

企業間関係を抽出する方法は，検索エンジンの結果を用いる．新しい企業リスト内のすべての企業名と入力企業を検索エンジンの機能である「AND」で結合し，クエリとして検索エンジンに投げかける．その際の企業名は，企業リストの生成で取得した正式名称・通称を含めて「OR」を用いて結合する．同様に入力企業も正式名称・通称を取得し,「OR」で結合したものを使用する.「AND」は論理積であり，リスト内の企業と入力企業の両方を含むページを検索エンジンに要求する．また,「OR」は論理和であり，この場合は正式名称で表示されていても，通称で表示されていても検索できるということになる．さらに，すべての名称は「""」で囲まれている.「""」で単語を囲むことにより，完全一致で検索することが可能である．

ここまでの例を示す．例えば,「日立製作所」と「ソニー」の正式名称・通称はそれぞれ「日立製作所」,「日立」,「HITACHI」,「株式会社日立製作所」,「日立グループ」と「ソニー」,「SONY」,「Sony」,「ソニー株式会社」,「東京通信工業」と取得できる．その場合,「日立製作所」と「ソニー」の関係情報を求めるために検索エンジンに投げかけるクエリは,「("日立製作所"OR"日立"OR"HITACHI"OR"株式会社日立製作所"OR"日立グループ")AND("ソニー"OR"SONY"OR"Sony"OR"ソニー株式会社"OR"東京通信工業")」となる．得られた検索結果の上位 L 件 (本実験では $L = 100$) のタイトルと概要文から検索クエリとして使用した単語を除いた文書集合を各企業の企業間関係情報 (*document*) として保存する．

3.3 クラスタリング

得られた *document* は，従来手法と同様に形態素解析機 MeCab を用いて形態素解析を行い，TFp 値と DFp 値をそれぞれ求める [徳永 99]．そして，ウォード法によるクラスター分析を行い，デンドログラムの作成を行う [永田 00]．

3.4 タグクラウドの生成

得られたタグクラウドに対して，従来手法と同様に

1. $TFp(t)$ と $DFp(t)$ 値を用いた評価値
2. 検索エンジンを用いた評価値

の 2 つの評価値を用いてタグを求める．タグクラウドとは，様々なタグを一挙に表示させたのもである．$TFCDF(t \mid C)$ によりフォントサイズを，$HIT(t \mid C)$ によりフォントカラーを示すこととする．

4. 実験

本手法の有用性を示すために Ruby を用いてシステムを構築し，実験を行った．*seedURL* は従来手法との比較を容易にするために，従来と同じ「日立製作所」の URL を使用した．

図 2 はクラスター分析結果のデンドログラムである．また,

*1 http://wikipedia-lab.org:8080/WikipediaThesaurusV2/

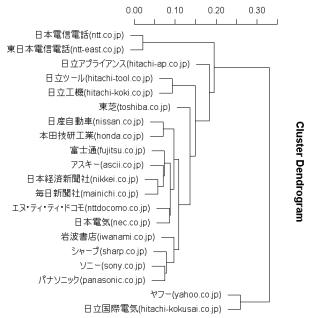

図 2: 提案手法：日立製作所のデンドログラム

タグ付けの結果に関して,「シャープ」,「ソニー」,「パナソニック」の 3 社の企業からなるクラスタ 7 のタグを図 3 に.「日立ツール」,「日立工機」の 2 社からなるるクラスタ 10 のタグを図 4 に,「日産自動車」,「本田技研工業」の 2 社からなるクラスタ 3 のタグを図 5 にそれぞれ示す．このクラスタ番号は，クラスター分析時にクラスタ化された順番に割り付けたものである．従来手法との比較が出来るように，従来手法のデンドログラムを図 6 に,「東日本電信電話」,「西日本電信電話」の 2 社からなるクラスタ 1 のタグを図 7 に,「日本ビクター」,「三洋電機」,「日本電気」,「東芝」の 4 社からなるクラスタ 8 のタグを図 8 に示す.

今回の提案手法が実際にクラスタリング精度の向上をもたらしたかについて，調査による評価を行った．調査手法は東京理科大学理工学部経営工学科に所属する学生にアンケート調査を行うものとした．調査票は従来手法と本研究の手法のクラスタリング結果のクラスタとタグクラウドをランダムで表示し，4 段階の評定法の回答形式を用いた．4 段階は点数の高いほうから順に「そう思う」（4 点）,「どちらかといえばそう思う」（3 点）,「どちらかといえばそうは思わない」（2 点）,「そうは思わない」（1 点）とした．表 1 に調査対象の属性を示した．また，アンケート結果を表 2 に示した．

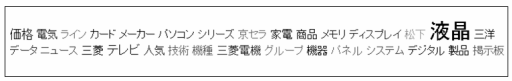

図 3: 提案手法：クラスタ 7 のタグ

図 4: 提案手法：クラスタ 10 のタグ

5. 考察

図 3 と図 6 のデンドログラムを比較すると，従来手法の方が企業数が多くなっていることが分かる．これは「企業リストに加える企業は URL の出現頻度の上位 20 件目と同等の企業までとする」という条件の下，企業リストを獲得しているからだと考えられる．

図 5: 提案手法：クラスタ 3 のタグ

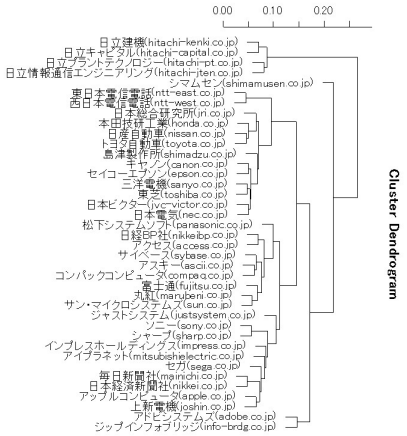

図 6: 従来手法：日立製作所のデンドログラム

表 1: 入力企業：アンケート調査:属性表

性別	男性	32 人
	女性	6 人
年齢	平均年齢	22.3 歳
就職活動経験の有無	有	29 人
	無	9 人

表 2: 入力企業：アンケート調査:結果

企業名	提案手法	従来手法
日立製作所	2.96	2.80
トヨタ自動車	2.47	2.90
JTB	3.09	3.04

日立製作所のタグを見てみる．クラスタ 7 のタグは「液晶」のみが目を引く表示になっている．これは，入力企業の日立製作所とクラスタに所属する「ソニー」,「シャープ」,「パナソニック」の 4 社が，液晶テレビのメーカであることを示せている結果である．クラスタ 10 のタグは，日立グループの工具メーカの 2 社が「工具」を目立つタグとしてクラスタリングされる分かりやすい結果となった．クラスタ 3 については，タグを見てもクラスタリングされた理由が分からなかった．そこで，クラスタするもととなったテキスト情報を見ると,「日立製作所」の正式名称・通称として取得した「日立」として検索されている関係情報が多かった.「日立」と自動車メーカーをクエリに並べて検索すると，中古車販売の日立店（茨城の地名）などの中古車販売のサイトのテキストを多数取得していたので図 5 のようなタグになったと考えられる．このような問題が起こるのは取得した正式名称・通称が，意図しない意味で, 他のクラスタの企業に関連の深いときに限られる．

「日立製作所」のアンケート調査の合計得点の平均値は提案手法の結果の方が高いという結果が出た．従来手法のクラスタ

図 7: 従来手法：クラスタ 1 のタグ

図 8: 従来手法：クラスタ 8 のタグ

リングの精度が低かった一因としては, 図 7，8 のように，表示させたタグクラウドのほとんどのタグが企業名であることが多かったためだと考えられる．企業名ばかりになる原因は就職人気ランキングのサイトや, 大学や研究室の内定先一覧のサイトを関係情報として取得してしまっていることである．そうしたサイトは企業間の関係情報を含まないことが多いので, 関係情報として取得されるべきではない．提案手法で述べたとおり，関係情報である *document* は企業リスト内の企業名を除いて使用する．提案手法の場合は，企業リスト内の企業名に対して正式名称・通称まで含めてすべて除いているので，就職人気ランキングのサイトや, 大学や研究室の内定先一覧のサイトからの関係情報をもとにタグを生成することは極端に少なくなったと言うことができる．

「トヨタ自動車」については従来手法のクラスタリング結果の方が平均値が高かった．主な原因としては,「富士通」と「トヨタ車体」のタグクラウドは，「全日本」,「女子」というタグが目立つ表示となっていて, 女子スポーツが強い企業同士であるが，そこまで詳しく知っている回答者は少ないと考えられるので平均値が低い結果となった．

6. まとめ

本論文では一つの企業を入力することで入力企業に関係のある企業リストを入手し，リスト内の企業を入力企業の正式名称・通称を取得しクラスタリングする手法を提案した．また，クラスタリングは入力企業との関係ごとにクラスタリングされていて，正式名称・通称の取得には Wikipedia を用いた類義語辞書を使用している．実験の結果多くの正式名称・通称を取得することが出来た．さらに, 取得できた正式名称・通称は関係情報を取得するだけでなく, *document* から除外する語としてとして, クラスタリング精度まで向上させることが分かった．

今後の課題としては, 類義語辞書として Wikipedia を使用すると, 間違った正式名称・通称を取得してしまうこと．さらに, 今回の日立製作所の実験結果で分かったように, トピック・ドリフト現象が起きてしまうことである．

参考文献

[takada 06] 高田 一樹: “検索エンジンを用いた企業間関係に基づく企業クラスタリング”, 東京理科大学卒業論文, (2006).

[nakayama 07a] N. Nakayama, T. Hara, S. Nishio: “A thesaurus construction method from large scale web dictionaries”, Proc.of AINA 2007.

[nakayama 07b] N. Nakayama, T. Hara, S. Nishio: “Wikipedia mining for an association web theaurus construction”, Proc. of WISE 2007.

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

[永田 00] 永田靖, 棟近雅彦: “多変量解析法入門”, サイエンス社 (2000).