Web情報を利用した関連企業のクラスタリングと可視化

Clustering and Visualization of affiliated company using Web information

前田 亮[*1] 松井 藤五郎[*2] 大和田 勇人[*2]
Ryo Maeda Tohgoroh Matsui Hayato Ohwada

[*1]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

[*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 recent years, for the number of companies with multiple business, to get the relationship between the companies which focus on a certain company have been difficult. In this paper, therefore, we propose a method to display the figure of relationship between companies with visual effects by using Web information. The method, first, we extract information from a certain company's affiliated company, then perform the clustering. Secondly, we display output result focused on a certain compay that can compare the strength of affiliated company.

1. はじめに

これまでは，ある分野の企業を知るためには業界地図などの紙媒体である書籍などを用いて調べられていたが，数多くの企業が存在する現在では，Web 上からそういった情報を調べることが一般的になっている．しかしながら近年，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いる．

近年，多くの企業は複数の事業を有しているため，就職四季報 Web などに存在するカテゴリサーチを用いて調べた場合，就職四季報などでは一つの企業につき一つのカテゴリのみでしか確認することができないという問題がある．

キーワードによって検索する方法でも多くは主要な事業のもののみしか記述されておらず，もし仮に全ての事業の内容を記述すると非常に膨大な量になるため，現実的には不可能である．また，上記の二点に共通する問題点として，ある企業をある特定のカテゴリに分類したり，その企業の事業領域を特定するためには，ある程度の専門的な知識が必要となる．すなわち，企業を対象に検索する場合，検索エンジンを用いると，ある企業を中心とした企業間の関係を求めることは困難である．

金らは企業間の関係のうち訴訟関係と提携関係に着目し，Web 上に公開されている情報から，企業間の関係を抽出する手法を提案した [1]．この手法は結果が視覚化されていて，高精度で結果が得られている．しかし，入力として，関係を求めたい企業リストや，どのような関係を求めるか (訴訟関係と提携関係等) などの関係情報を入力する必要があることや，企業間の関係が入力された関係に限定さえてしまうという問題点がある．

また高田らは，ある企業と他企業との企業間関係を検索エンジンを用いてクラスタリングするという手法について提案した [2]．この手法は，企業間関係ごとにクラスター分析をし，結果を出力するというものである．例として「日立製作所」についてのデンドログラムを図 1 に示す．しかしながら，出力結果がデンドログラムとして表示されており，このままではどの関連企業が入力企業と関係が強いのかというのが図 1 では読み取れない．

そこで本研究では視覚化ツールパッケージを用いて，入力企業に対する関連企業の強さを比較できる形でユーザが読み取りやすい出力結果を表示する手法を提案する．具体的には，調べたい企業の URL を入力し，その企業と関連企業との Google での完全一致検索におけるヒット数に着目して，入力企業に対する関連企業との関係図を作成する．

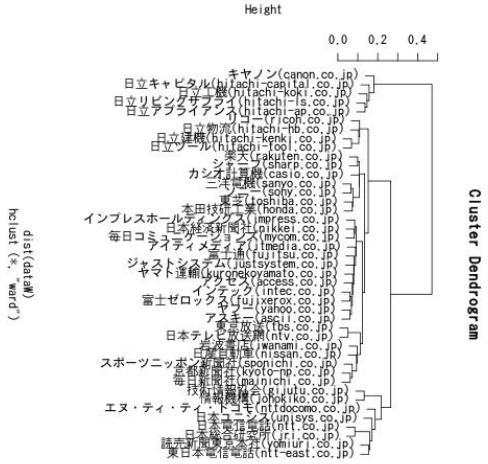

図 1: 高田らの実験結果

2. 提案手法

本研究の提案手法を以下に詳細に記載する．ただし，入力された企業の URL からその企業に関係のあると思われる企業の集合体である企業リストの生成は高田らの手法を用いる．また，本研究におけるシステム全体の流れを図 2 に示す．

連絡先: 前田 亮，東京理科大学大学院 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，j7409626@ed.noda.tus.ac.jp
なお 4/1 現在，松井の所属は [とうごろう機械学習研究所] に変更

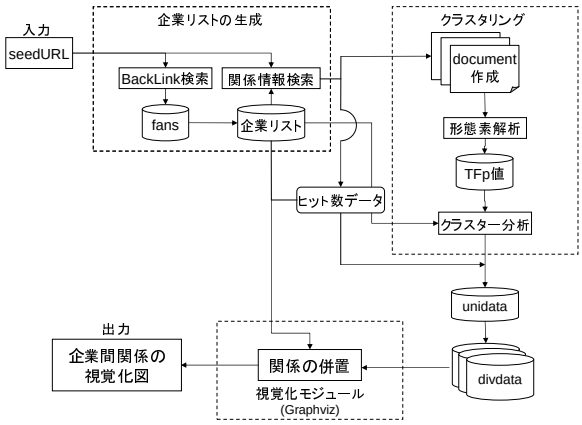

図 2: システム全体の流れ

2.1 クラスタリング

企業間関係情報を抽出する方法については，検索エンジン (Google) の結果を用いる．企業リスト内の全ての企業に対して，企業名と入力企業名の 2 つの企業名を並べてクエリとし，検索エンジンに入力する．ここで，検索エンジンの持つ機能で「""」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ため，本研究でのクエリは企業ごとに「""」で囲んだものを使用する．

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．

得られた *document* は形態素解析機 MeCab を用いて形態素解析を行い，*document* 内の名詞かつ，企業リスト内の企業名以外の単語のみを抽出する．その後，抽出された単語の *document* 内の相対頻度となる *TFp* 値 (1) を求める [3].

- 相対索引語頻度 $tf(t)$: ある文書 d 中に出現する索引語 t の数.

$$TFp(t) = \frac{tf(t,d)}{\sum_{s \in d} tf(s,d)} \quad (1)$$

得られた *TFp* 値を企業リスト内の各企業の要素ベクトルとし，クラスター分析を行う．ここではクラスタの境界を明確にするため，クラスター分析の非階層手法である k-means 法 (その他にも k-medoids 法などが存在する) を用いる．そして，本実験では企業リスト内の各企業を 8 つのクラスタに分ける．

得られた *TFp* 値を企業リスト内の各企業の要素ベクトルとし，クラスター分析を行う．ここではクラスタの境界を明確にするため，クラスター分析の非階層手法である k-means 法 (その他にも k-medoids 法などが存在する) を用いる．そして，本実験では企業リスト内の各企業を 8 つのクラスタに分ける．

ここで，クラスター分析には統計解析ソフト R を使用している．

2.2 企業関係の併置

3.1 における関係情報抽出より *document* とは別にヒット数データとして,「""」での完全一致によるヒット数のみを抽出したものを作成した．クラスター分析による結果の企業名 (企業の URL 付き)，クラスタ番号とヒット数データにおけるクラスター分析の各企業に対するヒット数を組み合わせたものを統合データ (*unidata*) として保存する．つまり，*unidata* には企業名，クラスタ番号，ヒット数が記載されている．そして，*unidata* 内に記載されているクラスタ番号ごとに *unidata* を分割し，この分割されたものを *divdata* として保存する．*divdata* には企業名，その企業のヒット数が含まれている．

次に，関連企業を *divdata* を用いて併置する．中心にフォントカラーを青とした *seedURL* の企業を置き，*divdata* が放射状に配置されるようにする．ここで，放射状に各企業を並べるために各クラスタごとにクラスタをまとめるためのフェイクポイントを置き，そこから各クラスタの企業を放射状に併置するように設定する．

また，各 *divdata* 内のうち入力企業に最も関係性が強い企業 (*divdata* 内で最もヒット数が大きい企業) のフォントカラーを赤となるようにする．

さらに，企業リストとヒット数データを組み合わせたものと *divdata* 内の企業名とを比較参照し，企業リスト全体を通して *seedURL* に関係が強いものほどフォントサイズが大きくなるようにする．

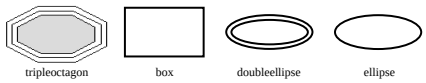

図 3: ノード形状

また，図 3 の 4 種類の形を使用して関連企業名が記載されたノード形状をヒット数によって変化させるようにする．ヒット数が 50 万件以上の企業を tripleoctagon，10 万件以上 50 万件未満を box，1 万件以上 10 万件未満を doubleellipse，1 万件未満を ellipse となるように設定する．

3. 実験

本提案手法の有効性を示すために Ruby を用いてシステムを構築した．そのさい，視覚化ツールパッケージである Graphviz[4] を用いて実験を行った．また，構築時に企業の URL を企業名に変換するため WHOIS サービスを用いた．このサイトは日本企業のみに対応しているため，企業リストの獲得には URL が.co.jp ドメインのみを使用している．本実験において *seedURL* は同業他社や，グループ会社が多く存在している「日立製作所」の URL(http://www.hitachi.co.jp) を使用し，企業リストを生成する際の backlink 検索取得件数を 50 件，企業リストの最低企業数を 50 件，関係文書取得件数を 100 件とした．図 4 は「日立製作所」を入力した際の出力結果である．

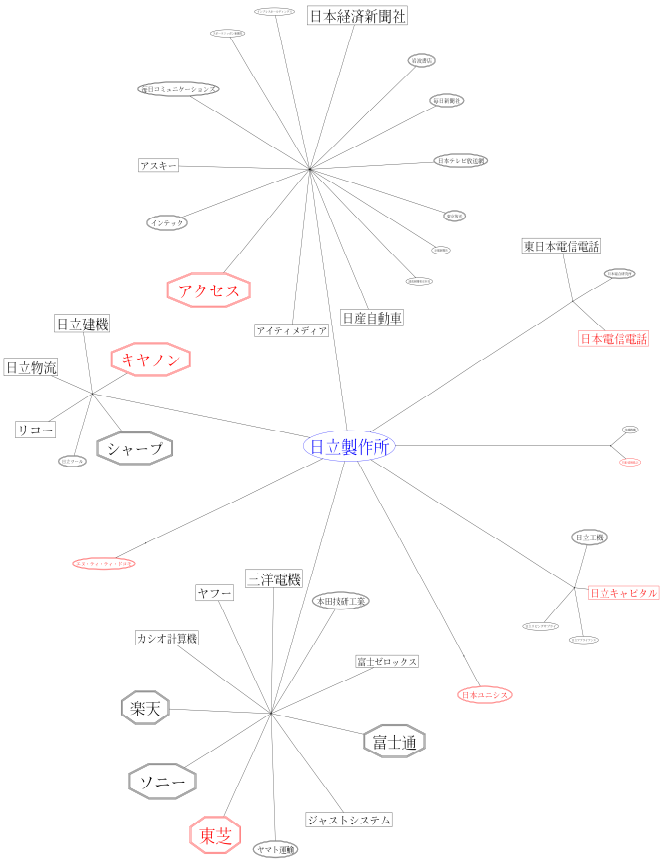

図 4: 出力結果 入力企業：日立製作所

4. 考察

クラスター分析において k-means 法を用いたため，類似性が高い関連企業間のまとまりが解りやすくなり，図 4 のように出力結果が得られた．さらに，中心に入力企業があり，また，キャノンが属するクラスタに注目した際，キャノンを中心にシャープや日立物流が配置されると図 5 のようになってしまう．そのため，日立物流等の企業がキャノンとの関係性があるように見えてしまうが，フェイクポイントを設置したことで，この関係図が日立製作所に関係していることが読み取れる．

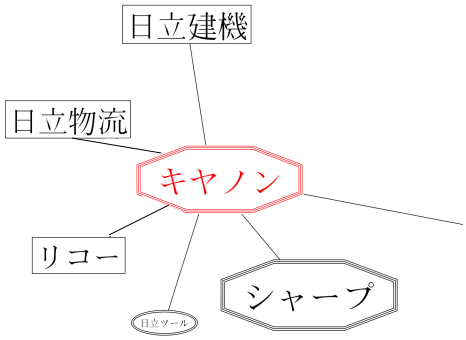

図 5: フェイクポイントがない場合の例

また，図 1 のようなデンドログラムは，取得された企業の中で類似している企業ごとに連結してしまっているため，入力企業との関係性の強さが見えなかった．しかし，前章での出力結果は入力企業との関係性をフォントサイズの変化で表したことで，入力企業に対する関連企業間全体の関係性が企業関係のフォントサイズが大きいほど強いということが見て取れる．図 4 での富士ゼロックスと三洋電機の関係がこれにあたる．三洋電機の方が富士ゼロックスよりもフォントサイズが大きいため富士ゼロックスの方が日立製作所と関係性が強いということが解る．

ただし，その一方で企業名が短い企業よりも長い企業の方が強調されてしまうこともある．これは，図 4 における東芝・ソニーの関係とアスキー・本田技研工業の関係に見られる．しかしながら，前者においては各クラスタ内において最も関係性がある企業名を赤字に設定していることで，また，後者においてはノードの形状をヒット数によって変更させることで，これらの問題点をクリアしている．よって，クラスタ内だけでなく，関連企業全体で入力企業との関係が比較しやすくなっていると言える．

5. まとめ

本論文では，調べたい企業の URL を入力することで，その企業に対する関連企業の強さを比較できる形で，ユーザが読み取りやすい出力結果を表示する手法を提案した．具体的には，入力企業と関連企業との検索システムでの完全一致によるヒット数に着目することで，ヒット数に比例して関連企業のフォントサイズを大きくした．また，一定のヒット数を超えた際にはノードの形状を変化させた．そしてこれらのことを，視覚ツールパッケージである Graphviz を用いてルール化し，企業間関係図を作成した．実験の結果，出力された企業間関係図は入力企業を中心に配置し，入力企業との企業間関係によって関連企業を様々な状態で配置することができた．このため，出力結果の表示を視覚的に理解しやすくすることができた．

今後の課題として，各クラスタがどのような特徴を持っているかを示せるように，関係情報からクラスタ其々の情報を抜き出せるようにすることで，さらにユーザにとって有益な出力結果にすることが可能になると考えられる．

参考文献

[1] 金英子，松尾豊，石塚満：Web 上の情報を用いた企業間関係の抽出，人工知能学会論文誌，22 巻 1 号 F，2007.

[2] 高田一樹，松井藤五郎，大和田勇人：検索エンジンを用いた企業間関係に基づくクラスタリング，東京理科大学理工学部経営工学科 卒業論文，2006.

[3] 徳永健伸：情報検索と言語処理，東京大学出版会，1999.

[4] AT&T 研究所：http://graphviz.org/ ，2007.