6-15

Webマイニングによる企業間関係のネットワーク抽出

日大生産工(院)　○後藤　大輔
日大生産工　　山下　安雄

1　まえがき

企業間には様々な関係がある。企業間の関係が織りなすネットワーク構造を分析することで、ある企業が他の企業と競争関係や提携関係であるかなど、企業の競争力の分析や取るべき戦略の決定に用いることができる。また、ネットワークの構築を分析することで、産業分野全体における、ネットワークの安定性、成長性も分析することが出来る。1970年以降、経済学や社会学の分野では、関係構造の背後にあるものを読み解く社会ネットワーク分析と呼ばれる研究が行われてきている。近年では、多種多様な情報がWeb上に公開され、Webから有用な知識を抽出しようとする研究が盛んに行われている。また、人間関係といった研究者ネットワークを抽出する研究や、国の関係、産業関係、企業間関係など、あらゆるレベルのネットワークを対象とした研究も行われている。本研究は、Web上に公開されている情報から、企業間関係レベルのネットワークを対象とし、企業間には、取引、提携、訴訟など、様々な関係があるが、特に提携関係に焦点を当てて、企業間の繋がりや強度の追求を行う。

2　目的

Webマイニングとは Webの内容、データ、文章から有用な知識、情報を発見するWebのデータを対象としたデータマイニングである。Web上には企業の関係を表す情報が入手可能であり、それらは2つに大きく分けられる。一つは、企業間の関係を整理してまとめてあるサイト、Yahoo!ニュースや日経のプレスリリースサイトなど整形された情報源は、Webに限らず会社四季報などの書籍からも企業間関係について知ることが可能である。もう一つは、企業が自ら公開するニュースリリースやWeb上のニュースサイトなど、分散した情報源である。検索エンジンを用いてWebページを集め、データマイニングを行い、Web全体の情報から企業間関係として、提携関係を抽出することを目的とする。

3　関係語の抽出

企業間関係が記述されたページを見つけるために、検索クエリに加える語を関係語と呼ぶ。本論文では、関係語を得る方法として、学習データを用いる方法と、Web上での共起を用いる方法を提案する。

3-1 学習データからの関係語取得法

Webマイニングを利用して提携関連を絞り込み、企業間の関係が含まれた多くのWebページを準備して、そのページに共通する単語を求める。業務提携や資本提携の企業間関係について書かれたWebページと、そうでないページを集め学習データを作る。各Webページから単語が出現するかしないかという属性を生成し分類を学習する。単語の出現を含むページの様々な特徴を属性とする分類問題になるが、検索エンジンに複雑なクエリを入力するのは難しいため、単語１語、もしくは複数個連言による組み合わせだけを調べる。学習の評価にF尺度を用いることにすると、この分類問題は、各単語が出現するかどうかによってF尺度がどう変化するかを調べればよいことになる。

F尺度は式(１)で定義される。

$$F_{Rel}(\omega)=\frac{2P_{Rel}(\omega)R_{Rel}(\omega)}{P_{Rel}(\omega)+R_{Rel}(\omega)} \quad (1)$$

ここで、$P_{Rel}(\omega)$は、単語又は連言を含むページのうち関係が正しく記述されたページの割合であり、$R_{Rel}(\omega)$は関係が記述されたページのうち単語ωが含まれるページの割合である。一つもしくは二つの関係語を用いて、検索クエリを生成し検索することで、より網羅的に関係を抽出することができる。し

Extracting Inter-business Relationship by Web Mining

Daisuke GOTOU, Yasuo YAMASHITA

たがって、F値が上位の複数の関係語を用いる。また、検索されたテキストの内容から企業間関係が実際に存在するかを判断するルールの中でも、この関係語を利用する。

3-2 Webを用いた関係語の取得法

関係語を得るために適切な学習データを用意するのは手間がかかる。そこで、本研究では、Webを用いて関係語を抽出することを考える。例えば提携関係を調べたいのであれば、「提携」という語を直接クエリに加えればよいというものである。さらに、「提携」とよく共起する語も提携関係を把握する手がかりになりそうである。そこで「提携」という語とよく共起する語をWeb上から獲得しようというものである。ここでは、「提携」などの語をω_{Rel}とする。そしてWeb上でのヒット件数を用いて、Jaccard係数、

$$\text{Jaccard係数} = \frac{|\omega_{Rel} \cap \omega|}{|\omega_{Rel} \cup \omega|} \qquad (2)$$

を計算し、これが高い語ωを関係語として用いる。ただし、$|\omega_{Rel} \cap \omega|$は「$\omega_{Rel}$連言ω」　をクエリにした場合のヒット件数、$|\omega_{Rel} \cup \omega|$は$\omega_{Rel}$選言ωをクエリにした場合のヒット件数である。なお、全ての語に対してこの値を計算するのは現実的でないので、ここでは企業間関係について書かれたページを用意し、そこから候補となる単語を切り出す。

4　企業間関係の抽出方法

システムの全体を図１に示す。前節に述べた方法で、あらかじめ提携関係の詳細（業務提携や資本提携）の関係語RWのリストを取得しておく。そして、システムに企業名のリストL（ｎ個）を入力しておき、企業（ｘとする）を取り出し、自分以外の企業（ｙとする）との関係の有無を調べてエッジを生成することで、企業間の関係のネットワークを出力する。全体の処理は、検索クエリの生成、関係ページの収集、関係の抽出という大きく３つに分けられる。検索クエリの生成フェーズ(MakeQ　ueries)では、関係語の上位$n_{queries}$個を企業名ｘ、ｙに加えることで検索クエリ集合Qを生成する。関係ページの収集フェーズ(DownloadP　ages)では、生成された検索クエリ集合Qを検索エンジンに入力し、上位にヒットしたページにn_{pages}件をダウンロードして関係の集合Dを得る。関係の抽出フェーズ(RelationExtract) では、ダウンロードしたページの内容を調べ、２つの企業に関する関係の記述があるかどうかを判断する。関係抽出フェーズでは、具体的に次の手順をとる。

１)収集した関係ページ集合Dに含まれる全ての文のリストSを収集する。

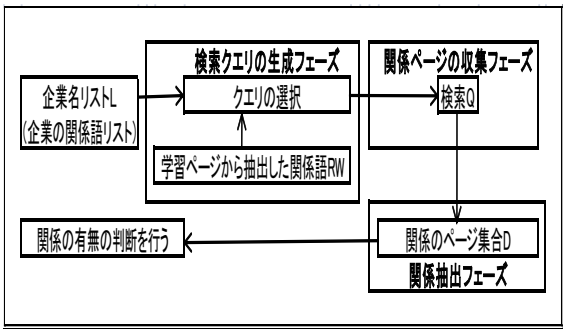

図１　システムの流れ

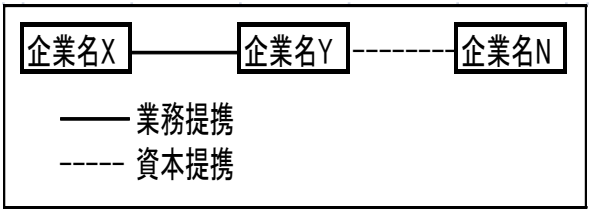

表１　企業間ネットワークの抽出例

２）２つの企業名と関係語ｒω（ｒω∈RW）が同時に出現する文ｓ（ｓ∈S）に対して、ｓに出現する関係語のスコアを足し合わせてその文のスコア$score_s$とし、全ての文の中で最もスコアの高いものをその企業間関係のスコア$score_{xy}$とする。

３)$score_{xy}$ が閾値$score_{thre}$を超えれば、２つの企業は関係があると判断する。この部分は構文解析や意味解析などより深い処理を行うことも可能であるが、本研究では出来るだけ簡単な方法にするため、このようなシンプルなルールを用いている。本研究では、$n_{queries}$＝２、n_{pages}＝５とした。なお、閾値は$score_{thre}$は、学習データからF尺度を使い、$score_{thre}$ と$score_{xy}$の差からネットワークの抽出を行う。

5　今後の展望

本研究の詳細関係は、業務提携と資本提携の２項目だったが、今回の結果を用い、詳細関係の項目を増やすと共に、今後も定期的に企業間関係を抽出することで、企業間の変化や動向の分析をしていきたいと考えている。

「参考文献」

1) 友部博教，松尾豊，武田英明，安田雪，橋田浩一，石塚満，Semantic Webのための人の社会ネットワーク抽出と利用，情報処理学会論文誌，Vol.46，No.6，(2005)，pp.1470-1479 .
2) 元田浩，津本周作，山口高平，沼尾正行，データマイニングの基礎，オーム社.
3) 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，Vol.22，No.1，(2007)，pp.48-57.
4) 松尾豊，中西紘子，橋田浩一，検索エンジンを用いた研究者の所属情報抽出，日本知能情報ファジィ学会誌，　Vol.19，No.6，　(2007)，pp.670-679 .