

D-001

Web 上の情報を用いた企業間関係の抽出

金英子† Yingzi Jin
 松尾豊‡ Yutaka Matsuo
 石塚満† Mitsuru Ishizuka

1. はじめに

企業の分析において、企業自身の資本金や従業員数などの属性を分析する方法もあるが、自分の企業が他の企業とどのような関係を持っているか、相手の企業が自分以外の企業とどのような関係を持っているかなどのような企業間の関係を分析することも重要である。本研究は、Web 上の情報を用いて、企業間の関係を抽出する手法について検討する。本論文では、企業間の負の関係である訴訟関係の抽出について述べる。これらの手法は、訴訟関係だけでなく、企業間の提携関係や合併・買収などの関係を抽出する時にも適応できる。

2. 問題定義

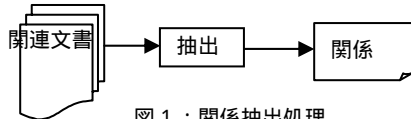


図1：関係抽出処理

関係抽出の研究では、図1のように、まず、関係を含んでいる関連文書を集めてきて、次に、それらのページを解析することで関係を定めることができる。例えば、Web 上の情報から研究者間の関係を抽出する研究[松尾 05]では、2人の研究者の名前を検索エンジンにクエリとして入力して、上位にヒットしたページを関連文書とする。そして、関連文書の中身から共著、共同開発などの関係を抽出している。しかし、企業間の関係を抽出する場合には、2つの企業の名前をクエリとして検索してヒットするページには、ノイズのページ¹がたくさん含まれており、また、上位のページだけではこの2つの企業間の関係を網羅的に抽出することができない²。Web 上の2つの企業の名前で共起するすべてのページを集めて関係を調べることも考えられるが、この手法は膨大なコストがかかるので実用的でない。上位のいくつかのページだけで、企業間の関係を抽出したいのが、本研究で提唱する考え方であり、そのためには、企業間関係のページを特定することのできる適切なクエリが必要になる。

3. 手法

3.1 システム概要

本研究のシステム概要図は図2の通りで、大きく分けて、検索クエリの生成、関連文書の収集、関係の抽出の三段階に分けられる。検索クエリの生成は、更に、関係語の生成段階と企業名ペアの生成段階に分けられる。

3.2 検索クエリの生成

(1) 関係語の生成

企業間関係のページを特定するための検索クエリとして、“企業1 企業2 関係語”(Q2)が考えられる。例えば、松下とジャストシステムの訴訟関係を調べたい時、

† 東京大学大学院 情報理工学系研究科 s

‡ 独立行政法人 産業技術総合研究所

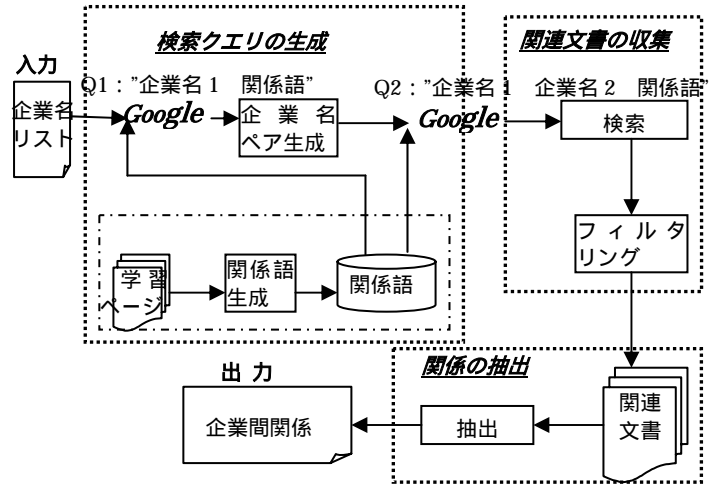


図2：システム概要図

“松下 ジャストシステム 訴訟”, 14,000件
 “松下 ジャストシステム 提訴”, 8,200件
 “松下 ジャストシステム 提訴 侵害”, 7,210件

などをクエリとして検索エンジン (google) に入力して上位のページを調べる。このような関係語を与えるために、我々は、予め学習ページの中から、関係を表しながら関係を特定できる関係語を用意しておく。本論文では、tf*idf による関係語の抽出と F 値計算による関係語の評価手法を提案する。tf*idf は、重要語を抽出する一般的な手法で、下記の式によって、学習ページの中から重要な単語を抽出する。

$$Weight(t) = tf(t,d) \times idf(t) \dots\dots\dots (1)$$

$$tf(t,d) = f(d,t) / f(d) \dots\dots(2), idf(t) = \log(N/df(t)) + 1.0 \dots(3)$$

また、重要度が高い上位の単語から、訴訟の関係語を選び、それらの関係語を組合せて関係語のペアを作る。

訴訟のページで重要な関係語がどれぐらい訴訟関係を特定できるかを評価するために、我々はテストページを集めてきて、関係語と関係語のペアの F 値を計算する。

$$F = 2PR / (P + R) \dots\dots\dots (4)$$

ここで、正解率 $P = cn / pn$ 、再現率 $R = cn / N$ である。ただし、 cn は、テストページの中で、関係語 (ペア) が現れる訴訟のページ数で、 pn は関係語が現れるすべてのページ数で、 N はすべての訴訟のページ数である。F 値が高いというのは、それらの単語 (ペア) が訴訟のページで頻りに現れて、正しく現れることを意味するので、訴訟のページを特定することができる。つまり、我々は、情報検索で評価のために使われる F 尺度評価手法を、検索クエリ生成段階で関係語を特定するための手がかりとして利用した。

(2) 企業名ペアの生成

企業名ペアを生成する時、最初に考えられるのがすべての企業名を組み合わせること (計算量は、 $O(n^2)$) である。しかし、この組み合わせは、企業の数が多くなるに連れて膨大になり、それに比べて、実際に関係のある企業ペアは

¹ 2つの企業の名前は出現するが2つの企業の関係でないページ。

² 企業間の1つの関係でもたくさんのニュースになりえる。

少ない。ここで、我々はそれぞれの企業の訴訟のページに現れる企業とその企業でペアを作る方法を提案する。つまり、“企業 1 関係語”をクエリ (Q1) として検索し、企業 1 の訴訟のページに現れる企業と企業 1 のペアを作る。これで、計算量は $O(n)$ に軽減できる。

3.3 関連文書の収集

(1) フィルタリング処理

クエリの生成段階で生成されたクエリ“企業 1 企業 2 関係語 (ペア)” (Q2) を検索エンジンに入力して検索すると、また、下記のようなノイズのページがヒットされると考えられる。つまり、企業 3 と企業 4 の訴訟のページに企業 1 と企業 2 が別の関連記事に現れる場合である。ポータルサイトの記事には特にこのような場合が多く、また、そのようなページはページランクが高いので、上位にヒットされる可能性が高い。ここで、我々は Simpson 係数を利用して、ヒット件数によって関係の強さを計算し、ある閾値以上の企業ペアだけに対して関連文書を収集する。

$$Rel_i(x,y) = Simpson(X, Y) \dots\dots\dots (5)$$

$$= \begin{cases} \frac{|X \cap Y|}{\min(|X|, |Y|)} & (\text{但し, } |X| > k, |Y| > k) \\ 0 & (\text{その他}) \end{cases}$$

ただし、 i は訴訟関係を表し、 $|X|$ と $|Y|$ は、それぞれの企業名と関係語をクエリ (Q1) として検索したときのヒット件数で、 $|X \cap Y|$ は企業名ペアと関係語をクエリ (Q2) として検索したときのヒット件数である。 k は $|X|$ と $|Y|$ のヒット件数の閾値を表す。Simpson 係数は、小さい企業からみた大きい企業との関係の強さを表すので、ヒット件数の差による影響を埋めることができる。

3.4 関係の抽出

これまでの処理では、訴訟関係のありそうな企業ペアに対して、できるだけ網羅的に訴訟関係のページとってきて関連文書とした。関連文書からの関係の抽出段階は、それらのページの中身をみて、本当にその 2 つの企業に関する訴訟関係が述べているかを確認することである。我々は、2 つの企業名と訴訟の関係語が近くに現れる文を抽出して、ヒューリスティックに、F 値の高い関係語 (ペア) を含む文を判断の基本となる重要文にした。

3.4 訴訟関係だけの独特性

(1) 訴訟の方向性

訴訟が他の関係と違うのは、訴訟にはどちらがどちらを訴えたかの方向性がある。我々は方向を示す助詞をクエリに追加することで訴訟の方向性を判断する手法を提案する。例えば、“松下が ジャストシステムを 訴え”で検索した場合に、61 件ヒットすることに対し、“松下を ジャストシステムが 訴え”は、17 件ヒットする。つまり、松下がジャストシステムを訴えたことが判断できる。

(2) 訴訟の段階の判断

既に和解になった訴訟段階は識別しておく必要がある。ただし、実際、訴えのニュースは和解のニュースより多いので、一般の訴訟の関係語でヒットした上位のページには、和解のページが含まれない可能性がある。ここで、我々は、関係語の生成段階で、学習のページを「訴えのページ」、「和解のページ」ように各段階に分類した。そして、それぞれの段階のページに対して、重要語を抽出し、ペアを作り、F 値計算を行った。ここで、F 値計算は、2 回必要になる。1 回目は、訴訟のページの中で、和解段階を特定するための訴訟ページ (学習ページ) の中での評価であり、

2 回目は、任意のページの中で、訴訟を特定するためのテストページの中での評価である。そして、訴訟関係が判断された企業ペアに対して、和解段階の関係語で再検索して、上位のページから和解になったかを確認すればいい。

4. 実験と結果

我々は、Yahoo ファイナンスに登録されている 313 個の電機機器の製造企業に対して、訴訟関係を調べた。

学習のページ¹から訴訟の訴えと和解の段階の関係語を抽出し、テストページ²からそれぞれの関係語 (ペア) の F 値を計算したところ、下記のような関係語と関係語のペアが上位に抽出された。

訴え段階の関係語 (ペア)	和解の段階の関係語 (ペア)
“侵害_提訴 // 提訴_地裁 // 提訴 // 侵害_提訴_地裁 // 侵害_提訴_求め // 和解_訴訟_侵害 // 和解_発表”	和解 // 和解_訴訟 // 和解_侵害

ここで、訴えの段階の関係語 (ペア) を利用して、企業名ペアを調べたところ、すべての組み合わせの 49,141 個のペアに対し、提案の手法では 1 回以上共起するペアは 11,402 個であった。次に、これらのペアに対して、方向性の判断と同時に、Simpson 係数による関係の強さを測り、閾値以上のペアに対して上位のページを関連文書としてダウンロードした。最後に、それらの関連文書から、企業名と関係語を含む文を抽出してヒューリスティックに関係のあるなしを再判断した。その結果、電機機器の製造企業の間には 2 個の訴訟関係があって、そのうち 1 個はすでに和解になっていた。この結果を、知財情報局のオンラインで提供する訴訟ニュースサイト (<http://news.braina.com/judge.html>) で、電機機器の製造企業間の訴訟ニュースを調べて比較した。

	我々の手法	訴訟ニュースサイト
訴訟関係の企業ペア	ミネベア 日本電産 (すでに和解) 日本電産 日本ビクター	ミネベア 日本電産 (すでに和解) なし

つまり、我々の手法で抽出した訴訟関係は正確で、訴訟記事をまとめているニュースサイトより網羅的であった。また、電気機器の製造企業の間には、訴訟関係が少ないということも分かった。

5. おわりに

本論文では、Web 上の情報を利用して、企業間の訴訟関係を抽出する手法について述べた。関連文書を集めるために、検索クエリに工夫を入れることで、検索エンジンを有効に活用した。更に、検索結果のヒット件数を企業間関係の強さを計量するための手がかりとして利用した。これらの手法は Web 上の情報が多くなるに連れて信頼性が高くなる。今後、これらの手法を利用して、企業間の提携関係なども抽出して、最終的には企業間関係のネットワークを作って、企業分析に役に立てる情報を提供したい。

参考文献

- [松尾 05] 松尾豊, 友部博教, 橋田浩一, 中島秀之, 石塚満: Web 上の情報から人間関係ネットワークの抽出, 人工知能学会論文誌 20 巻 1 号 E, (2005)

¹ 学習のページは、知財情報局の 2004 年 1 月 ~ 2005 年 3 月間のニュースから企業間訴訟に関する 130 ページにした。

² 任意のページは、任意のいくつかの企業名と上位のいくつかの関係語でそれぞれ検索して上位にヒットしたページをダウンロードして、手動で選んだ「訴訟のページ」150 個と、「訴訟でないページ」150 個である。