D-037

形式概念を用いた倒産情報の分析システム

Text mining of bankruptcy information using formal concept analysis

馬場 隆寛† 呉 小斌† 中藤 哲也‡ 廣川 佐千男‡

Takahiro Baba XiaBin Wu Tetsuya Nakatoh Sachio Hirokawa

1. はじめに

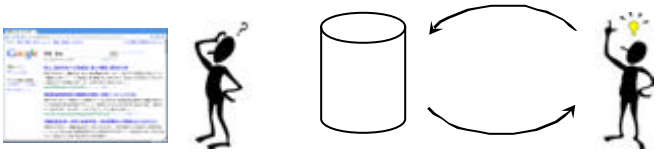
現在、web 上には倒産した様々な企業の情報などの社会の情報が存在し、容易に入手することができるようになっている。これらの情報から倒産した企業がなぜ倒産したのか、またどのくらいの企業が倒産しているのかといった情報が抽出できれば、企業の運営に役立てることができると考えられる。社会の情報を分析する研究として、金融市場の変動を分析する研究[4]や金融倒産企業をテキストマイニングにより分析する研究[5]が行われている。

本稿では、概念束を用いたテキストマイニングの研究[1]を応用し、倒産情報の文書に現れる単語の共起関係に概念束を適用し、内容を分析する手法を提案する。また、概念束におけるノードを特徴的なものに限定することで、簡潔な可視化を実現した。

2. 対話的・反復的分析システム

検索拡張として、図 1(a)のようなユーザが考えたキーワードに関連する単語を提示する検索システムがある。しかし、本稿で提案するシステムは一回の検索のヒントではなく、図 1(b)のように分析作業の途中で適切なヒントを提示し対話的・反復的な分析作業を支援するものである。分析を開始するきっかけは、ユーザが最初に決めなければならない。しかし、それ以降の分析における絞り込みや拡張の

方向性はシステムが提示する。ユーザは目的に従って選択することで、効率よく分析作業を進めることができる。関連語やヒントの自動的な提示によりユーザの知識では思いつかないようなきっかけが見つかることもある。



(a)検索エンジン　　(b)分析エンジン

図 1　検索エンジンと分析エンジン

3. 倒産情報とその基本的分析

本稿では、分析する倒産情報として、ウェブ上の情報を利用した。この倒産ニュースでは倒産した企業の倒産した理由が各企業について記されている。データは会社数が 726 社、文の数が 4799 文、一社当たりの平均が 6.6 個の文で書かれている。

倒産情報の基本的な分析として、まず地域×業種、地域×動詞のクロス集計と業種をキーワードとした特徴語の抽出を行った。分析に使用した単語は、地域、業種、動詞のそれぞれから出現頻度の高いものから選んだ。単語の横の()内の数はその単語を含む文の個数である。

表 1　業種×地域の分析

	東京(616)	大阪(209)	福岡(76)	北海道(73)
建設(306)	14	29	34	19
販売(281)	10	15	5	16
不動産(263)	4	11	4	8
製造(257)	5	7	1	4

表 1 は地域×業種のクロス分析を行ったものであり、地域として倒産している企業は東京が最も多いが建設関係の企業は大阪や福岡が東京よりも多い。このことから、都心部よりも地方の方が建築に関係する企業は、倒産しやすいことがわかる。

表 2　業種×動詞の分析

	申請(910)	倒産(863)	計上(704)	悪化(536)
建設(306)	46	16	14	18
販売(281)	75	71	35	98
不動産(263)	66	111	48	85
製造(257)	31	29	85	7

表 2 は地域×動詞のクロス分析を行ったものである。(不動産,倒産)が 111 件で最も多く、倒産している企業の業種は不動産に関わっているということが分かる。

表 3 は業種に対する特徴語を抽出したものである。建設に対して、土木、資材など関連しているキーワードが抽出できているが、倒産に関係するようなキーワードは抽出できていない。

†九州大学システム情報科学府　Graduate School of ISEE

‡九州大学情報基盤センター　Research Institute for IT

表 3　業種に対する特徴語

建設	土木 資材 建築 工事 公共
販売	住宅 業者 設立 用品 資材
不動産	市況 端 発する サブプライムローン
製造	部品 製品 電子 機械 装置

4. 概念束と包含グラフ

分類は対象を分析する最も基本的な手法である。どのような属性に着目して分類を行なうかにより得られる結果は異る。属性集合と対象集合が互いに特徴付けとなるような組を概念とよび、対象全体の分類階層を束として表現したものが概念束である[2,3]。本稿では、文書集合を対象とし、それらの文書に現れる単語を属性とする概念束を考える。概念束では、一つの概念（ノード）は文書集合と単語集合の組として捉えられ、互いに特徴づけあっている。また、隣接する概念により概念の分類を行っている。

図 2 は表 4 の行列に対する概念束である。表示の簡略化のため、それぞれの社名や単語は左端から右端に至るパスにおいて一度しか現れない。例えば、「D 社／客足」というノードに着目すると、「客足」はそれより右側の全てのノードに現れていて、「D 社」はそれより左側の全てのノードに現れている、と解釈される。従って、「D 社／客足」というノードは、「B 社、D 社」の二社が「バブル崩壊、客足」という二の単語で特徴づけられていることを示している。

このような簡略化表現を用いると、対象集合と属性集合がともに空となるノードが現れる。実際、対象集合と属性集合のそれぞれの要素数の大きい方を n とすると、概念の個数は、2^n となる可能性がある。しかし、簡略化表現だとノードに表示される対象名、属性名は高々2n 個しかない。従って、表示される概念束のほとんどのノードは空となり、可視化しても意味を捉えることができない。

そこで本稿では、概念束における空のノードを削除した包含グラフを提案する。さらに、ノードのラベルを単語だけとすることで、対象文書に現れる単語の関連を直観的に分析するできるようになる。図 3 は表 4 の行列より作成した包含グラフである。このグラフからは「不況」と「バブル崩壊」が「客足」を橋として関連していることが分かる。

表 4　簡単な行列の例

	A 社	B 社	C 社	D 社	E 社	F 社	G 社
リーマンショック	○		○		○	○	
不況			○				○
建築基準法			○		○		
客足				○			
バブル崩壊		○		○		○	

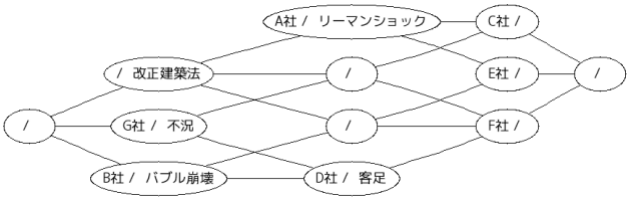


図 2　表 4 に対する概念束

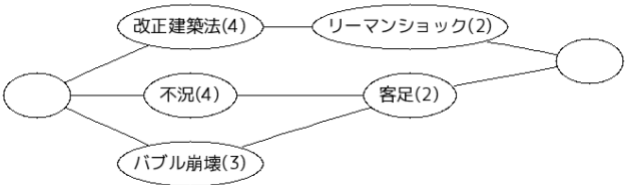


図 3　表 4 に対する包含グラフ

5. 包含グラフの概念束への埋め込み

文書集合 D、単語集合 W、および D,W の関係を表す $M \subseteq D \times W$ をまとめて(D,W,M)と書き文脈と呼ぶ。集合 D の要素 d、集合 W の要素 w について文書 d に単語 w を持っているとき$(d,w) \in M$ と表記する。$X \subseteq D, Y \subseteq W$ のとき(X,Y)が $doc(Y)=X$ かつ $word(X)=Y$ となるならば、(X,Y)を概念と呼ぶ。このとき(D,W,M)の概念全体を CL(D,W)と書く。ただし、$doc(Y)=\{d \in D \mid \forall w \in Y\ d$ は w を含む$\}$、$word(X)=\{w \in W \mid \forall d \in X\ d$ は w を含む$\}$。CL(D,W)の要素(X,Y),(X',Y')に対し、$Y >_{CL} Y' \Leftrightarrow doc(Y) \supseteq doc(Y')$と定義すると、順序集合$(CL(D,Y), >_{CL})$は束となり、これを概念束という。

二つの単語 $u,w \in W$ に対し、$u >_{HG} v \Leftrightarrow doc(u) \supseteq doc(v)$により定義される順序集合$(W, >_{HG})$を(D,W)の包含グラフと呼び、HG(D,W)と表す。

定理　包含グラフは概念束に埋め込むことができる。

すなわち、$\forall u,v \in W\ u >_{HG} v \Leftrightarrow f(u) >_{CL} f(v)$　という条件を満たす関数 $f:HG(D,W) \rightarrow CL(D,W)$が存在する。

証明

$w \in W$ に対して、$f(w) = (doc(w), word(doc(w)))$として上記が成り立つことを証明する。$u,v \in W$ に対し $u >_{HG} v$ であるとする。このとき、$doc(u) \supseteq doc(v)$ となる。ゆえに、$(doc(u), word(doc(v))) >_{CL} (doc(u), word(doc(v)))$であり、$f(u) >_{CL} f(v)$ である。逆も同様。■

6. 分析事例―改正建築基準法

「ダイエーファームと清水住宅の 2 社が自己破産」という文書から包含グラフを作成した。この文書に含まれる文は 7 個であり、包含グラフは図 4 のようになった。

企業の業務内容と倒産理由の 2 種類のことが書かれている。例えば、「住宅」、「基礎工事」、「マンション」などの業務に関係する内容と「受注」、「悪化」、「激減」

図 4　「工事」の包含グラフ

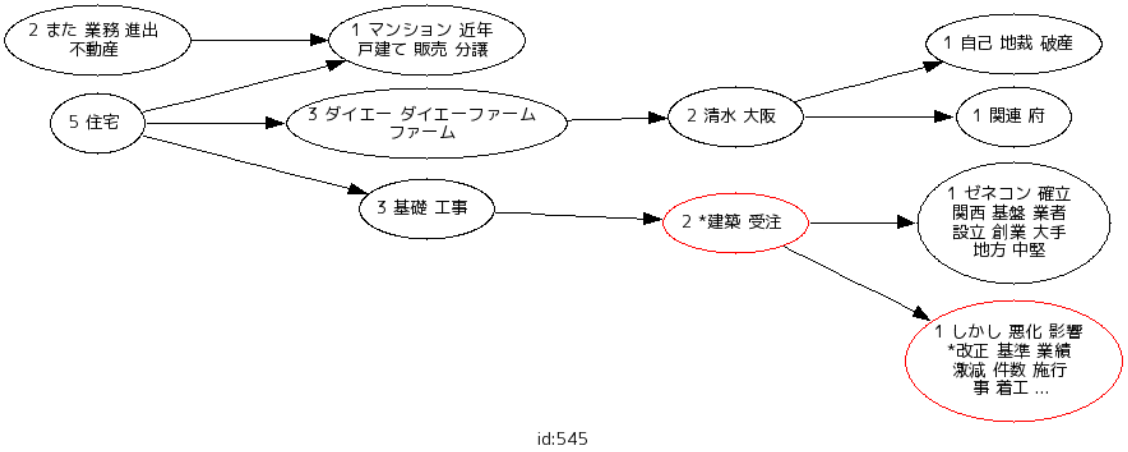
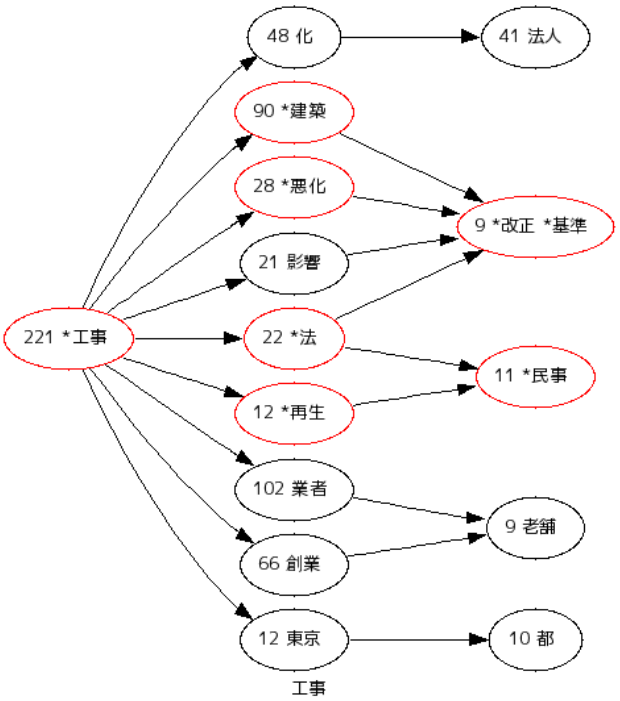


図 5　ダイエーファームの分析



などの倒産に関係する単語がみられる。「悪化」や「激減」という単語に関連する単語として、「建築」、「改正」といった単語が出現しており、改正建築基準法が影響しているのではないかと推測できる。実際にこの文書での倒産の原因は、改正建築基準法の施行の影響、不動産業務への進出にともなう借入金が財務を圧迫したとあり、グラフにそのことが出ていることが分かる。

ダイエーファームの倒産原因は、改正建築基準法であると推測したが、同じ分野の企業の倒産原因も改正建築基準法ではないかと考えた。そこで図 4 に現れる「工事」という単語に着目し、より一般的な分析を試みた。図 5 は「工事」についての包含グラフである。図 5 をみると「工事」に関連する単語として、「建築」、「改正」、「基準」という単語が出現しており、改正建築基準法が工事を行う企業の倒産原因となっていることが分かる。

図 6　「しかし　工事」の包含グラフ

図 4 の特徴として「しかし」という単語を含んでいるノードに正解の単語が含まれている。この例以外の場合でも「しかし」を含むノードに正解の単語も含まれていることが多かった。実際に文書を読むと、まず企業の業績や売上などの良い面を記した後に、逆接である「しかし」となぜ倒産したのかや業績悪化の原因が記されているためだと考えられる。

そこで「しかし」と「工事」を含む文に限定することで倒産要因を抽出できると考えた。図 6 は、「しかし　工事」をもとに作成した包含グラフである。図 5 よりもノード数がすくなくなり、倒産原因が改正建築基準法であることが分かりやすくなっている。

7. 分析事例―食品

「食品」という単語を含むものは 20 社あり、文は 45 個あった。この検索語から図 7 を作成した。

図 7 を見ると①販売と健康、②こめと安心、③冷凍と加ト吉、という 3 つの特徴に気づく。

「冷凍　加ト吉」について「食品」と「冷凍」で検索の絞り込みを行った結果、5 社、10 個の文が得られ図 8 のようなグラフとなった。加ト吉と取引のあるグループ企業についての倒産情報であることが分かった。

8. まとめと今後の課題

本稿では、倒産ニュースにある情報をもとに概念束に埋め込むことができる包含グラフを適用し、対話的かつ反復

的な分析システムを作成した。そのシステムにより倒産の原因の分析を行った。その結果、倒産の原因であるような特徴的な単語を抽出することができた。

包含グラフで気がつく単語の関連が倒産理由を必ずしも表すものではないという問題がある。それは、本稿で分析の対象とした倒産情報の文書が、企業概要の部分とその企業の倒産理由に関わる二つの異なる部分から構成されていることによる。例えば、図 8 で発見した「こめ」と「安心」のリンクは、食品についての事件や事故を想起させるが、この場合「おこめ安心食品」という企業名に由来するものであった。半構造的文書として取り扱うことで、このような問題の解決を今後検討する予定である。

また、包含グラフ有効性の定性的な評価方法や定量的な評価方法で検討しなければならない。グラフ中に含まれる赤いノードは倒産原因となると考えられる単語を含むノードであり、あらかじめ各文書を読み人手により決めている。これを用いて、定量的な評価を行うことができると考えている。

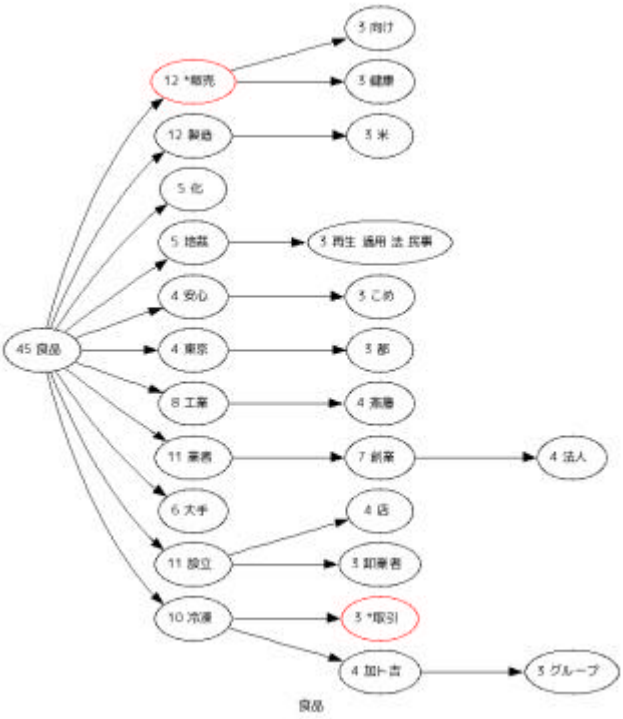


図 7　「食品」の包含グラフ

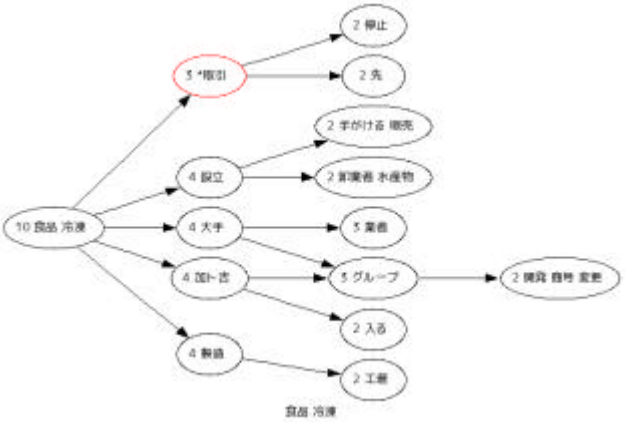


図 8　「食品」「冷凍」の包含グラフ

参考文献

[1]Takahiro Baba, Lucing Liu, Sachio Hirokawa Formal Concept Analysis of Medical Incident Reports, Proc. KES2010 (to appear)

[2]C. Carpineto, G. Romano, Concept Data Analysis Theory and Application, John Wiley and Sons, 2004

[3]B. Ganter, R. Wille, C. Franzke, Formal Concept Analysis Mathematical Foundation, Springer, 1999

[4]和泉潔,後藤卓,松井藤五郎, テキスト情報による金融市場変動の要因分析, 人工知能学会論文誌,Vol. 25, No. 3 pp.383--387, 2010

[5]竹内広宜, 荻野紫穂, 渡辺日出雄,白田佳子, テキストマイニングによる倒産企業分析, 経営情報学会 2008 年春季全国研究発表大会予稿集,2008