

プログラム・マネジメント支援のためのテキストからの 利害関係者の抽出と分類

Extraction and Classification of Stakeholder from Text to Support Program Management

吉田 守泰*¹ 吉岡 真治*²
Moriyasu Yoshida Masaharu Yoshioka

*¹北海道大学大学院 工学研究科 (現 日立製作所) *²北海道大学大学院 情報科学研究科
Hokkaido University (Now: Hitachi, Ltd.) Hokkaido University

Program management aims to support a management for a large and complex project. In this process, program managers should broaden their perspectives not only from a technological viewpoint but also from social, economical, cultural, and environmental views. Therefore, it is crucial to find out related stakeholders comprehensively for a good management. In this paper, we proposed a system that supports to identify stakeholders related to a program. In this system, we categorize stakeholders in two types. One is collaborator and the other is opponent. For extracting these two types of stakeholders, we collected typical expression for describing them and made pattern match rule for extraction. For evaluating our methodology, we implemented the prototype system that uses Japanese newspaper articles as a resource for extraction. We also conducted experiments to identify stakeholders.

1. はじめに

近年、企業には顧客の要求の多様化などを考慮した問題解決(製品の開発、工場建設など)が求められている。このような問題に対しては、多視点からの評価、分析が必要であるため、個人単独で問題解決にあたるのではなく、複数の選抜されたメンバーによる問題解決を行なうことが望ましい。このような複数のメンバーによって行なわれる問題解決業務をプロジェクトと呼ぶ。環境の変化の早い現在、ますますプロジェクトによる業務遂行が増えている。そこでプロジェクトを成功させ、成果を挙げるためには、プロジェクト全体の企画・推進・管理・調整を行う手法であるプロジェクト・マネジメントが必要になる[小原 02]。

さらに、最近では海外での工場建設や都市再開発事業など、複雑かつ大規模な課題に取り組むことも多く、そういった場合、1つの事業に対しプロジェクトがいくつも必要となる。いくつものプロジェクトが絡み合っている事業をプログラムというが、このような大規模なプログラムにおいて、個々のプロジェクトのマネジメントだけでは効率良く事業が進まず、目的を達成できない恐れがある。よって、事業を成功させるためには、プログラム全体を見渡し、複雑なプログラムを複数のプロジェクトに分解し、これらを統合的に協調させて管理することが必要となり、このプロジェクト管理手法をプログラム・マネジメントと呼ぶ[小原 02]。

プログラム・マネジメントを行なうためには、初期段階であるプログラム策定段階において、プログラムの全体像を把握し、個々のプロジェクトにおける業務目的・任務などを明確にする必要がある。しかし、一人の専門家が複雑かつ大規模なプログラムの全体を把握することは非常に困難である。そこで新聞記事や Web、自社のデータなどからキーワード検索などにより、過去にプログラム事例について問題や話題となった記事を提示できれば、プログラム事例に関する理解の手助けになるのではと考えられる。

しかし、キーワード検索には

- プログラムについて把握できていない状況を想定しているため、適切なキーワードを与えることが難しく、検索結果を絞り込むことが困難である。
- 検索結果をテキストとしてのみ提示するため、閲覧性が悪く、結果の確認がユーザにとって負担となる。

といった欠点が挙げられる。

そこで新聞記事などのテキストからステークホルダー(利害関係者)とその間の関係などを自動で抽出し、図1のようなステークホルダーの関係図を新聞記事とともに提示することができれば、プログラム事例の理解の手助けになるのではないかと考えられる。よって本論文では、情報抽出分野の技術[関根 04]を利用し、ステークホルダーを自動で抽出・分類するシステムを提案する。

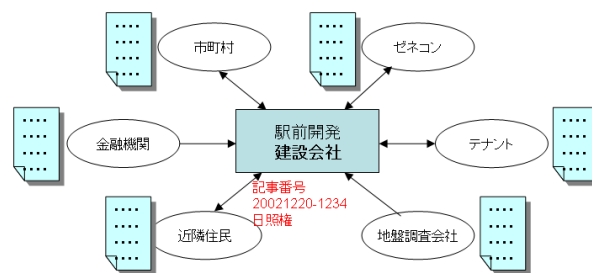


図 1: ステークホルダーの関連図

2. テキストからのステークホルダー情報の抽出と分類

2.1 テキストからの情報抽出

近年、大規模文書を情報源として、必要な情報を抽出し整理する手法である情報抽出の研究が広く行なわれている[関根 04]。例えば、[小林 03]らは、評価表現に特徴的な表現に関する情報を用いて、Web上に存在する文書から様々な商品に関する

連絡先: 吉岡真治, 北海道大学大学院情報科学研究科, 札幌市北区北 14 条西 9 丁目, 011-706-7107, yoshioka@ist.hokudai.ac.jp

評価の情報を抜き出すための方法を提案している。具体的には、評価表現に特徴的な表現を係り受け構造を考慮したパターンを作成し、そのパターンに当てはまる文書から、評価表現（例えば、製品とそれに対する評価）の組み合わせを抽出する。これらの手法は、全ての情報を漏れなく取り出すことができるわけではないが、大規模な文書の内容から役に立つ情報のみに圧縮できるという点で有用であると考えられる。

2.2 プログラム・マネジメントに役立つステークホルダーの抽出と分類

本研究では、先へのべたパターンを用いた情報抽出の技術に応用し、プログラム・マネジメント支援のためのステークホルダー情報の抽出を行なう。一般に、何かのプログラム（プロジェクト）に関わるステークホルダーの情報を見つけるためのパターンを作成するのは困難であるが、本研究では、次の2つの場合を想定することにより、パターンの作成を行なった。

- プログラム（プロジェクト）で利害関係者との関係が悪化した場合には、訴訟と言う形で表面化する。
- プログラム（プロジェクト）で利害関係者との間に協調的な関係を築き、共同で作業を行なう場合には、プレスリリースと言う形で新聞報道がなされる場合がある。

この仮説に基づき、本研究では、訴訟と共同作業のプレスリリースにまつわる新聞記事を対象として、パターンを作成し、ステークホルダー情報の抽出を行なう方法を提案する。

また、用いたパターンの種類に応じて、ステークホルダーとの間に協力関係があるのか、対立関係があるのかを分類する。また、実際の新聞記事には、具体的なインスタンスとしてのステークホルダーが多数現れるため、ステークホルダー間の関係が複雑なネットワークになる。これを解消するために、本研究では、プログラムに対応するステークホルダーのカテゴリーを設定し、具体的なインスタンスであるステークホルダーのカテゴリーへの分類を行なう。

次に、ステークホルダー情報抽出システムを実現するのに必要となる要素について考える。

プログラムと関連しない記事から情報抽出を行なっても、適切な情報が得られないため、適切な情報抽出を行なうためには、適切な情報検索システム（検索部）が必要である。この記事群に対し、パターンを用いた情報抽出を行なう抽出部が必要である。さらに、抽出したステークホルダーをいくつかのグループへ分類を行なう分類部を用いて、情報を整理することにより、ユーザにとって理解しやすいステークホルダーの関係の表示を行なうことが可能となる。図2に、新聞記事データベースからの一貫したステークホルダー情報の抽出と分類を行なうための、作業の流れを示す。

3. パターンを用いたステークホルダー情報の抽出

3.1 パターンの作成と利用

本研究では、訴訟と協力関係のプレスリリースを対象としたパターンの作成を行なう。具体的には、「AがBにCをめぐる～を求めた訴訟」という文章からは、AというステークホルダーとBというステークホルダーの間にCという事柄に関して問題が起きたという事が読み取れる。そこで、「(主体1)が(主体2)に(対立内容)をめぐる～を求めた訴訟」というパターンを作成することによって、新聞記事からの情報抽出が可能となる。

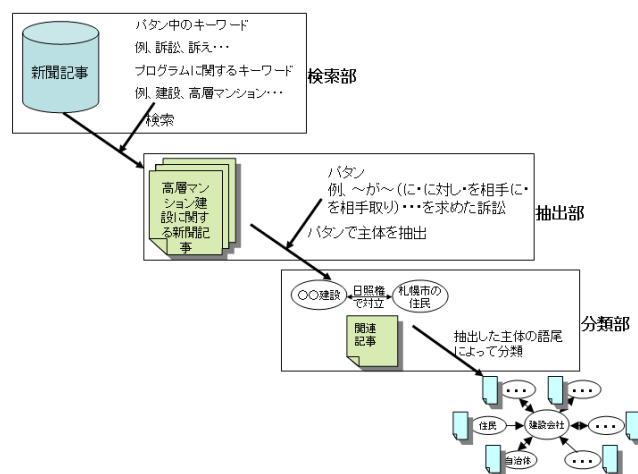


図2: ステークホルダー情報抽出の流れ

また、文字の出現順序をそのまま扱っていると「(主体2)に(主体1)が(対立内容)をめぐる～を求めた訴訟」といった語順の異なる文ごとに対応関係を記述する必要がある。また、単純に語順で行なうと、長い文のときに「～が」に相当する語が複数出てくる場合などがでてくるため、誤った語を抽出してしまう可能性が高くなってしまふ。よって構文解析の結果を利用する。

しかし、実際に構文解析システムとして Cabocha[工藤 02] を利用し、予備実験を行ったところ、訴訟に関するパターンについては入力文の係り受けの誤りによってほとんどのステークホルダーを抽出することができなかった。理由として、訴訟の文章は一般に長い文章が多く、誤った構文解析を行なってしまうことが挙げられる。予備実験を進める中で、例えば「…として、…が…に…を求める訴訟で…」という文章が存在した場合、「…として」の部分に「～が」「～に」という言葉を含むことは多いが、それ以降に含むことが少ない傾向があることがわかった。よって、誤抽出の可能性は増えるが、網羅的な情報抽出という観点から、訴訟に関しては「～が」「～に」という抽出する部分に関しては係り受けを考慮せず、パターン中で最も後方のものを抽出することにした。

3.2 作成したパターン

今回、日本経済新聞(2001年度分)を対象に図3に示すパターンを作成した。訴訟に関するパターンと提携・協力に関するパターンを作成するために、訴訟と提携・協力に関するキーワードで検索した新聞記事それぞれ100記事程度からパターンとなるような文の形を探した。実際には何度も繰り返し出てくる形や、主体の関連性が強い形などをパターンとして作成した。パターン1から5までは「訴訟」に関するパターン、パターン6から9は「提携・協力」に関するパターンである。A、Bにあたる部分をステークホルダーとして抽出し、Cにあたる部分をその間の関係として抽出する。

パターン1から4は、「～に」、「～に対し」、「～を相手に」、「～を相手取り」といった4種類の形がパターンとして考えられたので、それらのいずれかがあった場合に主体として抽出する。また、訴訟に関するパターン1～5に関しては、主体間の関係を抽出するために、パターンにあてはまった文章の中で「～を巡(めぐ)って」、「～を巡(めぐ)り」、「～に絡んで」、「～として」といった部分があった場合、その部分を抽出する。

パターン7と8の「…提携・協力」の部分は「提携する」、

- パターン 1 A が B (に・に対し・を相手に・を相手取り) … を求めた訴訟
- パターン 2 A が B (に・に対し・を相手に・を相手取り) … を求める仮処分
- パターン 3 A が B (に・に対し・を相手に・を相手取り) … を求める訴え
- パターン 4 A が B (に・に対し・を相手に・を相手取り) … を求めた住民訴訟
- パターン 5 A が損害賠償請求した訴訟
- パターン 6 A は B と共同で ~ をする
- パターン 7 A は B と (C で・C でも) (… 提携・協力) することで合意した
- パターン 8 A は B と (C で・C でも) (… 提携・協力) (した・する)
- パターン 9 A は B と (C で・C でも) 提携 (すると・したと) 発表した

図 3: 利用したパターン

「協力する」という語の他に、「業務提携した」、「包括提携した」などといった語もあるので、そのような語もカバーできるように設定した。

また、ステークホルダーとなりえるものの品詞は名詞、未知語が考えられる。それ以外にはカッコが主体の語尾になっていることも多い。それらを抽出するために品詞が「記号」となっているものも抽出をする。また「～など」という表現も多く存在するので、品詞が「助詞-副助詞」となっているものも抽出することにする。

3.3 抽出の単位

抽出した語はステークホルダーの関連図として表示するため、意味の分かる単位で抽出できていることが望まれる。また、抽出した語を「企業」、「自治体」などのいくつかのグループに分類するのだが、そのためにはある程度長い単位で主体を抽出しておく必要がある。そこで主体も主体間の関係も両方も、基本的には初めに見つけた語から係り受けのリンクを上に通っていきけるすべての語をつなげたものを主体として抽出する。しかし、主体については係り受けのリンクをたどっている途中に、「～として」「～」、「～は」「～が」という語にたどり着いたときは、そこで意味が切れている可能性が高いため、そこから上には辿らないこととする。

4. プログラムの関連記事の検索

プログラムに関するステークホルダーを抽出するためには、事前にプログラムに関する新聞記事を準備しておく必要がある。本システムではパターンが適用可能な文書からのみ情報抽出が行なえるため、パターン中に存在する語のいずれかを含んだ文書が情報抽出の対象となる文書となる。よって、このパターン中に存在する語とプログラム事例に関するキーワードを必ず含む新聞記事が、今回行なうステークホルダーの抽出に役立つ記事と考え、情報検索を行なった。この情報検索には、本研究室で開発されたブーリアン検索システムと確率型情報検索システムを組み合わせた検索システム ABRIR を用いた [Yoshioka 04]。

5. ステークホルダーの分類

パターンを用いた抽出によって、プログラムに関連する 2 つの主体と、その間の関係が抽出できていることになるが、このままではステークホルダーの組がばらばらにたくさん収集できただけで、ステークホルダーの関連図にはなっていない。そこで、抽出したステークホルダーをいくつかのグループに分類する必要がある。

5.1 分類方法

本研究では、固有表現抽出 NEXt [渡辺 04] を参考にしたステークホルダーの分類を行なった。具体的には、設定した分類辞書の語が抽出した語の語尾にあるかどうかでグループ分けをする。例えば、もし建設会社のグループの分類辞書に「建設」という語があり、「大手建設会社の建設」といった語が抽出されたとする。この場合、抽出したステークホルダーの語尾に建設会社の分類辞書に存在する「建設」という語が含まれているので、建設会社のグループに分類する。

また、抽出した語の語尾が「～など」「カッコ」「～ら」で終わっていることも多いので、「辞書に設定した語+(など・()ら)」という形でもそれぞれのグループに分類を行う。

また、今回は建設に関するプログラムを対象としているので、建設会社のグループはもれなく分類を行ないたい。そこで、建設会社のグループについては、語尾が「…(など・社)」となっているステークホルダーでさらにその中に「建設」「ゼネコン」といった語を含む場合も建設会社のグループに分類する。

5.2 作成した分類辞書

今回は図 4 に示す「建設会社」、「企業」、「自治体」、「団体」、「住民」、「個人」という 6 つのグループに対して分類辞書を作成した。

建設会社	建設、建設会社、建、建物、建設工業
企業	社、会社、地所、銀行、経営者、証券、工業、商事、不動産、セメント、不動産業者など 39 語
自治体	都、道、府、県、市、区、町など 11 語
団体	組合、団体、会、連合
住民	住民、世帯、住民側
個人	人、男性、女性

図 4: グループごとの分類辞書

6. 実験

作成したシステムの有効性を検証するための評価実験を行なった。図 2 に示した情報抽出の流れを踏まえた上で、各段階での評価を行なう。

6.1 実験方法

実験は「マンション建設」というプログラムを想定し、日本経済新聞 (2002 年) を対象に行なった。まず本システムの検索部においてプログラムに関するキーワードである「マンション」、「建設」という語の両方を必ず含み、かつパターン中の語である「共同」、「協力」、「提携」、「訴訟」、「訴え」、「仮処分」の中のどれかを含むという条件で検索を行なう。また、前述し

た 9 つのパターンと分類辞書を用いてそれぞれ抽出・分類を行なう。

6.2 実験結果

表 1 は検索性能に関する実験結果である。検索部のみでの出力は 90 記事見つかった。この 90 記事に対して抽出部により情報抽出を行なった結果、15 記事から情報の抽出が行なわれた。

表 1: 実験結果

	検索部のみ	検索部 + 抽出部
適合記事数	43	11
不適合記事数	47	4
合計記事数	90	15

まず検索部のみでの出力についての精度は 47.8% (43/90) となった。次に抽出部を組み合わせた出力についての精度は 73.3% (11/15) となった。2 つを比較すると抽出部を組み合わせた方が検索部のみでの出力結果より、精度に関しては良い結果が得られたことがわかる。また抽出部での出力について再現率は、25.6%(11/43) となり、あまりいい結果は得られなかった。

6.3 考察

実験において抽出部を組み合わせることによって、検索精度は向上した。よって、今回作成したパターンが適切なステークホルダーを抽出することができる良いパターンであるということが言える。しかし、検索部で検索した適合記事のうち 25% 程度しか抽出部において出力することができなかった。これは今回作成したパターンの数では不足していることが考えられる。より多くの適合記事を抽出するためにはさらにパターンを追加する必要があるが、取りこぼした記事の中にパターンとして作成できそうなものも存在した。例えば、「～は～に対し～を求める」といった形や、「～は～と組んで」、「～は～に対して～を要請した」といった形は、主体間の関係が強く、また他の記事においても出現するのではないかと推測できる。

次に、不適合文書に関する考察を行なう。不適合記事を抽出してしまった原因の 1 つに、主体の抽出に関しては上手くいっているが、検索部においてマンション建設に関係の無い記事を検索してしまっているものがある。これに関しては抽出・分類部で対応することは難しく、検索システムの性能向上や検索式の工夫など検索部に頼るしかないと考えられる。

もう 1 つはステークホルダーとしてふさわしくない語を抽出してしまうことによって誤った記事を出力してしまったものである。本来、パターンとしてあてはまるべきではない文章にあてはまってしまい、例えば「販売」という語や、「同システム」という語を抽出してしまい、不適合記事を出力してしまった。これに関しては両方ともパターン 6 によって抽出してしまったものであるため、パターンとして適切であるかどうかを考える必要がある。

6.4 抽出・分類の考察

次に抽出・分類における性能について考察する。まず、抽出についてはステークホルダーの抽出精度は 88.2%(30/34)、関係については 77.8%(7/9) となり、良い精度を示した。分類についても抽出された語 34 語のうち、26 語 (76.5%) に対し正

しい分類をすることができた。しかし誤分類や分類されなかった語もあるため、分類辞書の修正も必要である。

7. まとめ

本研究においてプログラムマネジメント支援を目的としたテキストからのステークホルダーの抽出・分類手法を提案した。実験において、検索性能や抽出・分類の精度に関してよい結果を示すことができた。またプログラム事例の理解に役立ちそうな記事やステークホルダーを出力できたのではないかと考えられる。しかし、検索・抽出・分類において漏れ・誤りも多く見られる。

今後は、様々なテキストに対し実験を行ない、パターンの追加・修正、あるいは分類辞書の修正をすることにより、さらに性能を向上させることが必要であると考えられる。

謝辞

本研究の一部は IMS プログラム「プログラム・マネジメント支援システムの研究開発」により、実施された。また、本システムについてコメントを頂いた、当プログラムのメンバーである清水建設(株)、住友電工(株)、都立大学の皆様方に感謝の意を表す。また、日経全文記事データベース 日経 4 紙 DVD-ROM 版 2001 年版 2002 年版を利用した。

参考文献

- [小林 03] 小林, 乾, 松本, 立石, 福島: テキストマイニングによる評価表現の収集, 情報処理学会自然言語処理研究会, 2002-NL-154, pp. 77-84 (2003).
- [関根 04] 関根聡: 情報抽出 - 情報を整理して提示する -, 情報処理, Vol. 45, No. 6, pp. 563-568 (2004).
- [渡辺 04] 渡辺, 榊井, 福本: 固有表現抽出ツール NExT の精緻化とユーザビリティの向上, 言語処理学会第 10 回年次大会発表論文集, pp. 413-415 (2004).
- [Yoshioka 04] Yoshioka, M. and Haraguchi, M.: Study on the Combination of Probabilistic and Boolean IR Models for WWW Documents Retrieval, in *Working Notes of the Fourth NTCIR Workshop Meeting, Supplement Volume*, pp. 9-16 (2004), <http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/WEB/NTCIR4WN-WEB-YoshiokaM.pdf>.
- [工藤 02] 工藤, 松本: チャンキングの段階適用による日本語係り受け解析, Vol. 43, No. 6, pp. 1834-1842 (2002).
- [小原 02] 小原重信: P2M 入門, 株式会社エイチアンドアイ (2002).