3B1-2

プロジェクト進捗報告書の表層表現を用いたトラブル予測

村上 明子 *1*2　　中村 大賀 *1
Akiko Murakami　　Taiga Nakamura

*1日本 IBM 東京基礎研究所
IBM Research, Tokyo Research Laboratory

*2東京大学大学院 学際情報学府
Interdisciplinary Information Studies, The University of Tokyo

ソフトウェア開発などのプロジェクト進捗管理として、リーダーがプロジェクトの状況を管理者に定期的に報告し、管理者はその情報を元に評価を行うという方法が一般的である。しかし管理者は多数のプロジェクトを同時に管理する必要があるため、プロジェクトの状態を簡便に把握するために通常はスケジュールやコスト等に関する定量的な数値のみに注目することが多い。プロジェクト管理においてトラブルをできる限り早く認識することは非常に重要であるが、報告書上の数値に変化が現れるのはしばしばトラブル発生後であるためトラブルの早期発見には適さない。そこで、熟練した管理者は報告書に書かれたテキストを手がかりとしてトラブルが起きそうなプロジェクトを発見して監視することによりトラブル予測を行っている。本論文は表層表現の出現頻度とトラブル発生の関係を用いたトラブル発生予測能力を示すイベント予測特性の指標を提案した。また、実際のデータを用いて実験を行い、結果からイベント予測特性指標の補正を行い、改善を確認した。

1. はじめに

ソフトウェア開発などのプロジェクトを管理する立場において、計画の遅延やコストの超過などのトラブルをできる限り早く認識することは非常に重要である。トラブルの発見が遅れるほど、状況が複雑化して対策が困難になったり、余分な労力やコストがかかるリスクが高まったりするためである。しかし、トラブルは複合的な理由で起こることが多く、トラブル発生の予兆を感知することは必ずしも容易でない。典型的なプロジェクト進捗管理では、プロジェクトのリーダーが情報を定期的に上層部の管理者に報告し、管理者はその情報をもとにプロジェクト状態を評価し、過去の経験や他のプロジェクトとの比較から、そのプロジェクトが健全な状態であるかを確認する。

定量的な値によって明らかなスケジュールの遅延、コストの超過などが報告されれば、そのプロジェクトが危険な状態であると認識をして対策を講じることになる。しかし最初に述べたようにトラブルの認識は早いほど有利であり、そのような定量的な値に変化が現れる前にトラブルの予兆を感知することが望まれる。熟練したプロジェクト管理者であれば、数値データだけではなく報告書のテキストの内容や表現の特徴などから、今後当該プロジェクトにトラブルが発生するかどうか予兆を感じとることができると言われているが、全ての管理者がテキストを用いた予測に熟練しているわけではない。また多数のプロジェクトを管理する必要がある場合には、管理者が全ての報告のテキストを時間をかけて精査することは困難である。

本論文はテキストを分析し、トラブルの予兆である表現を発見することを目的とする。そのために、各プロジェクトにおいて生成された定期的な報告書を時系列のデータとして扱い、管理者がトラブルを認識したことをイベントとみなすことにより、時系列データが複数あるときにテキスト表現によるイベントの予兆を発見するモデルを定義する。本論文が対象とする通常の時系列データとイベントの関係は通常のモデルとは異なり、プロジェクトにおけるトラブルはプロジェクト固有のイベントである。そこで、我々は時系列のテキストを、イベントの生起した時刻を基点として整列し、イベント発生までの時刻（イベントの後であれば負の時刻）をテキストの時刻情報として扱う。

本論文では、トラブル弁別に有用な表現の抽出のために、テキスト内の表層表現が出現したときにプロジェクトがトラブルとなる確率を求め、表層表現のトラブル発見能力の指標として用いることを提案する。また、その指標を表現のプロジェクトにおける分布に基づいて改良し、その効果についても議論を行う。

2. 関連研究

業務の日報や故障の報告書などのテキストを解析し、他の業務や故障解析に役立てる研究は数多く行われてきた [1, 2, 3, 4]。

市村ら [1] は営業管理職が営業日報から成功・失敗事例を分析するため、要因概念と結果概念をテキストから抽出し、営業活動における要因と結果の因果関係を示した。また、安藤ら [2, 3] は船舶の故障報告書を分析するため、テキストから故障に関するイベント抽出し、抽出されたイベントの関係を分類した。いずれの研究も業務知識を有する人によって作成されたオントロジーを用いてルールベースで抽出され、相関ルールや含まれる単語の概念上の距離などから関係性を求めている。

宅間ら [4] の研究はイベントをオントロジーを用いて抽出した後、報告書を時系列的に扱っている。彼らはコールセンターのコールログを、コールの発生した時刻の情報を元に時系列データとして扱い、イベントの増加を初期に発見し、製品の不具合が多く報告される前に発見するというものである。

このような時系列的な文書からの特徴発見の研究もいくつか行われている。Kleinberg [5] や藤木 [6] らは、何らかのイベントが起こったとき、そのイベントに関する文書（blog, 新聞記事など）が多く発生するという仮定を用い、ある話題が注目された期間を抽出している。Kleinberg [5] は文書の作成される時間間隔が定常状態において一定であると仮定しているのに対し、 藤木 [6] らの手法では掲示板の書き込みなどのように、時刻によって文書発生数が変化しているときに対応する拡張を行っている。

連絡先: 村上明子, akikom@jp.ibm.com

3. 時系列データとイベントとの関係

我々は、プロジェクトが通常とは異なるトラブル状態（スケジュール遅延など）に陥る前に予兆したいという目的のために、プロジェクトの進捗管理のために作成された報告書に着目する。定期的に作成される報告書の作成日時と内容を取得すれば、時系列の情報を含んだ文書情報を得ることが可能である。

2 節で示したように、従来研究では時系列の中で急激に増加した事象や、外部イベントの生起を示す表現の抽出などが行われてきた。しかし、これまで対象となっていた時系列文書は文書全体が同じ時系列を持っており、その時系列の中でのイベントと表層表現との関係を見ていた。例えば宅間ら [4] の場合は、製品のトラブルという外部イベントを、多くの時系列を持ったテキストから発見するという問題を扱っていた。これに対し、本論文で扱う時系列データは各プロジェクトから定期的に作成される報告書であり、そのプロジェクトがトラブル状態と認識されるという事象がイベントとなる。従って、各イベントはそのプロジェクトに対してのみ生起する。そこで、イベントが発生したプロジェクトの文書をイベント発生時点を基点とした時系列文書として扱い、イベントが発生しなかったプロジェクトの文書は表現の現れやすさの指標を求めるために用いる。次章で、このイベント予測能力を持った表現を抽出するための手法について述べる。

4. 表現のイベント予測表現抽出

この節では、熟練したプロジェクト管理者が知見として持っている「トラブルの予兆と思われる」表現を、データを元にして自動的に抽出することを示す。このトラブル（イベント）予測をする能力のある表現を、我々はイベント予測表現と呼ぶ。経験の豊かな管理者ならば「こんな表現を報告書で読んだあと、そのプロジェクトがトラブルとなることが多い」といった直感を持っているといわれる。イベント予測表現の性質を定量的に測るため、過去データにあるイベントのある／なしと表現の頻度の情報を元にし、イベント予測能力の指標を求め、その値を元にイベント予測表現の抽出を行う。

4.1 候補表現

表現のイベント予測能力を調べるために、候補となる表現を各報告書から抽出する。この論文において、文書中の表現とは名詞、接続詞などの単語や、名詞と用言といった係り受けのパターンなどを指す。

本論文で用いた候補表現の種類を表 1 に示す。これらを表現を抽出するため、あらかじめ文書内のテキストに対して形態素解析および構文解析を行った。単語は抽出対象の品詞である場合に表現として扱った。

また、イベントの兆候は名詞などの単語だけではなく、「何がどうした」というような名詞と用言の係り受けで表される表現も含まれると考えられるため、単語のほかに係り受け関係を持つ名詞と用言についても表現の一種として扱った。係り受けは、構文木の中から係り受け可能な名詞と用言の組を取り出すことにより抽出した。単語および係り受けに現れる活用語尾は終止形に集約したが、否定についてのみ区別をおこなった。

さらに、格助詞の中には状況を限定する役割を持つ場合があるため、係り受けに関して格助詞の区別をつけたものも候補として抽出した。例えば、「○○が終了した」と「○○は終了した」を区別することで、後者の表現が「終了していないもの」の存在を示唆する可能性を捉えることができる。

表 1: トラブル予測表現候補語の単語およびパターン

単語	名詞、動詞、形容詞、形容動詞
係り受け	名詞→述語、名詞→格助詞→述語

4.2 イベント予測特性値の付与手法

ある表現 e がイベント発生に関連しているかどうかは、イベントが発生したプロジェクトに属する文書に含まれている割合で測ることができる。我々の目的はイベントの発生前に、イベントが発生するかどうかを予測することにある。ここである表現がイベントの予測能力を持つということは、表現 e があるイベントが起きたプロジェクト内に含まれ、かつ観測された後にイベントが生起していることであると定義する。そこで全体の文書集合を D 、イベント発生前の文書集合を D_{prev} とし、文書 d に含まれる表現 e の頻度を $f_d(e)$ とすると、表現 e のイベント予測能力を示すイベント予測特性値 $Cap(e)$ は以下で定義される。

$$Cap(e) = \frac{\sum_{d \in D_{\mathrm{prev}}} f_d(e)}{\sum_{d \in D} f_d(e)}$$

これは、イベントが起きる前の表現 e の頻度を、文書全体の e の頻度で割ったものと等しい。もし表現 e がイベントの発生したプロジェクトにおいてイベント発生前のみ観測される場合、$Cap(e)$ は最大値 1 をとる。逆に e の出現が完全にランダムである場合、総和の期待値は全体のイベント発生率になる。つまり、 $Cap(e)$ が 1 に近づくほど e はイベント予測に有用であることになる。

5. 実験

5.1 実験方法

前節の議論に基づき、外部イベントとして「プロジェクトのトラブル」を予測するための表現を実プロジェクトのデータを用いて調べた。各プロジェクトで作られた報告書の内容と作成日時、およびプロジェクトの状態に問題があると認定された場合にはその記録日時を取得した。これ以外にも、イベント発生の定義として具体的な数値データ（コストやスケジュールなど）が特定の状態に陥った時点を使用することも可能である。

4.2 節で述べた手法を用いて、文書から抽出した各表現に対してイベントの予測しやすさの指標であるイベント予測特性値 $Cap(e)$ を求める。候補のパターン（品詞、係り受け）によるイベント予測特性の傾向のの違いを見るため、品詞および係り受けパターンごとにイベント予測特性値を求めた。

5.2 結果と考察

今回の分析対象のデータでは、形容詞、形容動詞は頻度が 3 以下の表現しか存在しなかった。頻度が低いとたまたまトラブル前の報告書に出現した可能性もあるが、頻度が低いながらもイベント予測特性値の高い中には「センシティブだ」「敏感だ」や、「大事だ」「核心的だ」といった同義の表現が多数見受けられた。このことから、同義の表現を集約すれば、イベントの予測表現として扱えるのではないかと考えられる。

名詞、動詞と係り受けに関しては、頻度がある程度あり、かつ予測特性値の大きいものが見られた。指標の大きさの順位とイベント予測特性値の関係を図 1 に示す。図 1 には、名詞

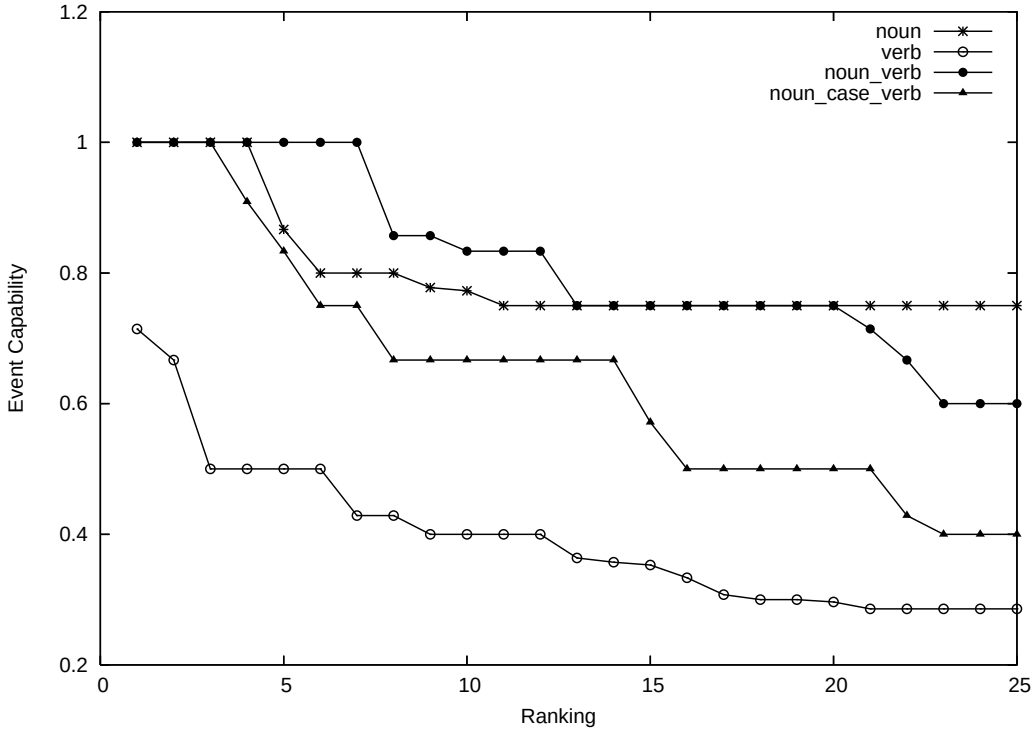

図 1: イベント予測特性値と特性値順位の関係

(noun)、動詞 (verb)、名詞と用言の係り受け (noun_verb)、名詞と用言の係り受けを名詞が係る格助詞で区別したもの (noun_case_verb)、計 4 種類の結果を示した[*1]。図 1 より、名詞と用言の係り受けのトラブル予測特性値の上位は 1 に近いものが多いのに対し、動詞は上位でも特性値が低い結果がみられる。これより、名詞や名詞と用言の係り受けの方がトラブルを予測する能力が高い表現が多く含まれていると推測される。一方、格助詞まで含めた係り受けは異表記が多くなり各々のトラブル予測特性値を下げてしまうという結果がみられる。

名詞については、上位に高い特性値を持ったものが多数存在した。上位に現れた名詞の詳細を調べると、直感的にトラブルと関連しそうな表現もみられる一方で、プロジェクト固有の表現と思われるものも多くみられた。例えば、トラブル予測特性値の値が 1 である 4 つの表現のうち 3 つはプロジェクトを特定できるような表現であった。このように、前節に述べたモデルではプロジェクト固有の表現や報告書の書き手であるプロジェクトのリーダーがたまたま偏って用いる表現についても特性値が高くなる可能性がある。このような表現を取り除くため、我々はトラブル予測特性値を補正することを試みた。

6. イベント予測特性値の補正案

実験で示されたように、4.2 節で提案した $Cap(e)$ はプロジェクト固有の表現や特定の人が多用する表現に対しても大きくなってしまう可能性がある。この問題を緩和するため、イベント予測特性値の補正を行う。上記の考察に基づき、表現のイベント予測能力は複数のプロジェクトに普遍的に現れる表現ほど高いと仮定した。従って、表現が複数プロジェクトに分散している割合をイベント予想特性に反映させた。イベントが発生したプロジェクト数を N_{event}、イベント前に表現 e が出現したプロジェクト数を $N_{\mathrm{prev}}(e)$ とすると、プロジェクト分散を加味したイベント予測特性 $Cap_{\mathrm{mod}}(e)$ は以下のように定義される。

$$Cap_{\mathrm{mod}}(e) = Cap(e) \times \frac{N_{\mathrm{prev}}(e)}{\min\left(\sum_{d \in D_{\mathrm{prev}}} f_d(e), N_{\mathrm{event}}\right)}$$

$Cap_{\mathrm{mod}}(e)$ の第 2 項の分子はイベント前に表現 e が出現しているプロジェクト数、分母はイベントが起きる前の表現 e の頻度とイベントが生起したプロジェクト数とを比較し、小さい方の値をとったものである。この項はイベント前に現れた表現 e がどれくらい異なるプロジェクトに分散しているかを表しており、全てのプロジェクトに表現が出現している (表現 e の出現頻度が高い場合) か、表現の全てが異なったプロジェクトに現れる (出現頻度がイベント生起プロジェクト数より小さい場合) ときに最大値 1 をとる。また、表現が一つのプロジェクトに偏って出ているときに最小（表現 e のイベント生起前における頻度の逆数）となる。

この $Cap_{\mathrm{mod}}(e)$ を用いて、再度表現のイベント予測特性値を求めた。名詞において頻度 4 以上、イベント予測特性値の上位 25 個を取得して人手で判断したところ、補正前に上位 25 表現中 10 表現あったプロジェクト固有の表現が、補正後には 2 表現まで減少することが観察された。

この補正は表現のプロジェクトにおける分散度合いによって値を低めるだけで、値は上昇しない ($Cap(e) \geq Cap_{\mathrm{mod}}(e)$)。そのため、格助詞を考慮した係り受けについては値は小さいままである。しかし順位には変化が見られた。上位に現れた 25 表現のうち、限定の意味をもつ格助詞「は」によって係り受け

*1 頻度 4 以上の表現で比較した

関係になっている表現は、補正前の 3 表現から 7 表現へ増加した。補正により抽出できた例は「定義は終了する」「品質は問題ない」といった、問題を一部のみ報告しているような表現であり、これは問題を一部しか報告していない、あるいは認識していないプロジェクトがトラブルとなりやすいという管理者の直感と一致していると考えられる。

用言、係り受けに関しても同様に、出現の傾向が変化した。ここでは表現がプロジェクトにどれだけ分散しているかについて考慮したが、複数プロジェクト文書を同じ人が作成していた場合には表現の属人性の影響を扱うためにさらに別の補正が必要である可能性もある。

7. 各表現のイベント発生の時系列確率分布

今回、我々はトラブル予測に有用な個々の表現の抽出を試みた。最終的には抽出されたこれらの表現を用いて、文書が入力された際のトラブル予測を行うことが目的である。そのため、ある表現が出現した際のトラブルの時系列的生起確率分布を求める手法を提案する。これは、文書の出現後に「トラブルになる/ならない」を判定するだけではなく、時期的な情報も加味するものである。

表現 e における、その表現の発生した時刻を基点としたイベント発生の頻度分布を求めることを考える。そのために、文書集合 D_{prev} をプロジェクトごとのイベント発生時刻に対する相対時刻によって分割する。文書が作成される日時は離散的であるため、一定の期間内に作成された文書をまとめて扱うために時間区間を定める。そのため、イベント発生から時刻 τ 前の時刻のの前後 $(\Delta t/2)$ 内に作成された文書を同時期に作成された文書として扱い、この文書集合を $D_{\mathrm{prev}}(\tau)$ とする。すなわち、ある文書 d の作成時刻を基準とするイベント発生の相対時刻が、$\tau - \frac{\Delta t}{2}$ と $\tau + \frac{\Delta t}{2}$ の間にあるとき、 $d \in D_{\mathrm{prev}}(\tau)$ であるとする。すると、表現 e が発生した時刻 τ 後のイベント発生確率 $\Pr(\tau, e)$ は以下で定義される。

$$\Pr(\tau, e) = \frac{\sum_{d \in D_{\mathrm{prev}}(\tau)} f_d(e)}{\sum_{d \in D_{\mathrm{all}}} f_d(e)}$$

$\Pr(\tau, e)$ の分母は表現 e の文書全体での頻度、分子はイベントが発生したプロジェクトにおけるイベント発生から τ 前の文書における表現 e の頻度の合計である。これを各表現ごとに作成し、イベント特性リスト $P(e)$ とする。

こうして作られる特性リスト $P(e)$ は、表現 e の観測を起点としたイベント発生の頻度分布を示している。これらは各表現の時系列的なイベント発生の確率を示しているが、文書が入力であるときには各表現の共起なども影響するため、文書に対するイベント生起確率を求めるにはさらなる定式化が必要である。今後の課題として、これらのイベント予測表現と特性リストを用いて、文書が入力された際のイベントの発生確率分布を求めることが挙げられる。

8. まとめ

本稿では複数のプロジェクトの時系列的な文書である報告書から、イベントであるトラブルの認識を予測するために有効な表現を抽出する手法を提案した。はじめに、トラブルが起こる前に作成された文書集合と全体の集合それぞれに出現する表現の頻度の比によって、その表現のトラブル予測特性を付与した。実験により、あらかじめ抽出しておく表現の種類により傾向に違いがあることが観察された。また、プロジェクトや人に固有な表現が多く見られるという傾向から、プロジェクトに特異な表現を排除するために表現のプロジェクトにおける分散を考慮した補正を提案し、有効性を確認した。

また、イベント予測特性の高い表現に対して時系列的なイベント発生確率分布を得る手法についても提案を行った。今後の課題として、今回得られたイベント予測表現の集合とこれらの時系列的イベント発生確率分布を用いて、文書全体を入力とするイベント予測モデルの構築が可能であると考えている。

謝辞

本研究に対して、細川宣啓氏、渡辺千恵子氏に多数の助言を頂きました。深く感謝を致します。

参考文献

[1] 市村, 中山, 赤羽, 三好, 関口, 藤原：“日報分析システムの開発”, 電子情報通信学会技術研究報告. NLC Vol.100 No.401 (2000).

[2] 安藤, 大和, 堀, 増田, 白山：“テキストマイニングを用いた故障報告書分析手法の研究”, 日本造船学会論文集, **192**, pp. 475–283 (2002).

[3] 安藤, 大和：“テキストマイニングによる船舶故障データの分析 (<特集>製造現場における信頼性)”, 日本信頼性学会誌 : 信頼性, **26**, 8, pp. 906–912 (2004).

[4] 宅間, 野美山：“テキストデータを用いた問題の早期発見手法”, IPSJ SIG NL-162 (2004).

[5] J. Kleinberg: “Bursty and hierarchical structure in streams”, Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2002).

[6] 藤木, 南野, 鈴木, 奥村：“document stream における burst の発見”, IPSJ SIG NL-160 (2004).