

Twitter におけるユーザープロフィールと拡散力の関係分析

プロジェクトマネジメントコース 矢吹研究室 1142016 井上 乃祐

1. 序論

Twitter は 2006 年に開始したサービスで、コミュニケーションツールのひとつとして利用されているソーシャル・ネットワーキング・サービスである。国内ユーザーは 2014 年 6 月では 1980 万人、2015 年 5 月では 2390 万人であり、月間アクティブ率は 70.2 % である [1]。

Twitter はツイートと呼ばれる 140 字以内の短い文字列を投稿するサービスである。自分以外のユーザーのツイートを読むためには、そのユーザーのページにアクセスする方法以外に、そのユーザーをフォローすることでツイートを読むことができる。フォローしているユーザーのツイートはまとめられ、タイムラインを形成する。

そのほかの機能にリツイートと呼ばれるものがあり、それは他の人のツイートを再びツイートするというものである。自分の画面上に流れてきたツイートをリツイートすると、自分のフォロワーの画面にも流れる。同じように、自分がフォローしているユーザーがリツイートすれば、自分のタイムライン上にリツイートが流れてくる。リツイートされるツイートには、ツイート内容という情報以外にアイコンや、ユーザーの ID などの本質以外の情報も含まれる。

Twitter を利用していると、フォローしているユーザーからリツイートが流れてくることがある。その流れてきたリツイートを見てみると、似たような内容でもリツイートされた回数に違いがあることに気づいた。そこで私はプロフィールのアイコンがリツイート率に影響があるのではないかと考えた。

2. 目的

Twitter のアイコンが拡散力に影響があるかを調べ、情報の本質でないアイコン部分が本質に与える影響を調べる。

3. 手法

VirtualBox をインストールして仮想マシンを作成し、そこに Ubuntu をインストールする。その後、Ubuntu 上で Python と Tweepy をインストールして、Streaming API を使用するプログラムを実行し、リツイートされたツイートを集める。

その後、リツイートされたツイートのアイコンをいくつかのタグで分類し、R を用いて重回帰分析をする。



図 1 収集したデータ

4. 結果

若い男性、中年の男性、子供の男の子のタグをつけたアイコンが多くリツイートされている。

5. 考察

上記の結果を見ると、若い男性、中年の男性、子供の男の子のタグをつけたアイコンが多くリツイートされていた。そこから男性のアイコンのほうが、ツイートに信頼感や、説得力があるからではないかと考えた。

6. 結論

Twitter では男性のアイコンのほうがリツイートされる割合が高いと言う結論がでた。

参考文献

- [1] GaiaX SocialMedia Lab. 【保存版】3分で総復習！2015年、11のソーシャルメディア最新動向データまとめ.
[http://gaiax-socialmedialab.jp/socialmedia/435\(2016/1/14 閲覧\).](http://gaiax-socialmedialab.jp/socialmedia/435(2016/1/14%20閲覧).)

ビッグデータ処理技術を用いた Wikipedia マイニング

プロジェクトマネジメントコース 矢吹研究室 1242005 石井康之

1. 序論

Wikipedia は、多くの人がボランティアで執筆するオンライン百科事典プロジェクトである。

Wikipedia は 2001 年 1 月 15 日に創設され、2001 年 5 月ごろに日本語版が発足した。

多くの人が参加するプロジェクトの代表例である Wikipedia を調査することによって、このような形式のプロジェクトのマネジメントについての有意義な知見が得られることが期待できる。

このオープンなオンライン百科事典プロジェクトの成功理由について様々な考察がされており、その中の 1 つの要因に「適切な時に、それぞれのニッチに対しボトムアップとトップダウンの適切な混合率を出していた」というのがある [1]。Wikipedia では、ボトムアップの力、無統制の善良さで成長したと見えるのだが、実際にプロセスをよく調べてみると、見かけ以上に入念なトップダウンのしくみによる管理がなされている。

Wikipedia の編集者の中には、管理者と呼ばれる利用者たちがおり、この利用者たちが統制を行っているのではないかと考えた。

そこで当研究では、管理者の動向を見るため、管理者の編集回数がどのように変化しているか調査する。結果から Wikipedia の成功理由に、この管理者がどのように関係しているか見つけ出す。

2. 目的

Wikipedia を一つのプロジェクトとみなし、このオンライン百科事典で管理者の動向がどのように変化しているか調査する。

3. 手法

以下のとおり手法を行う。

1. Wikipedia 日本語版の編集履歴まで含んだファイルをダウンロードし、ローカルでデータマイニングを行う。
2. Wikipedia の管理者の編集回数の変化を解析する。

3. 管理者の編集の割合がプロジェクトの動きにどのようにつながっているのか調査する。

4. 結果

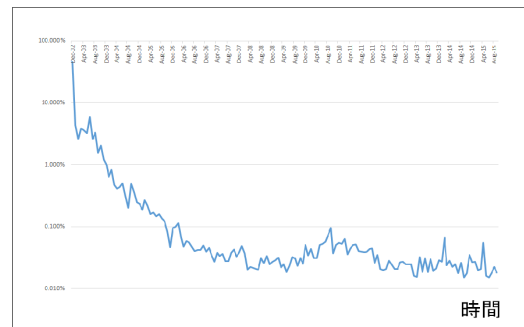


図 1 1ヶ月ごとの総編集回数における管理者の編集の割合の変化

日本語版 Wikipedia が実装されてからの、管理者の月別編集回数のデータ解析を行った。1ヶ月ごとの総編集回数における管理者の編集の割合を出した。

5. 考察

日本語版 Wikipedia では、管理者が編集を行う必要性が薄れてきている。実装当初から、管理者は毎月約 300~400 回ほどの編集を行っていたが、2015 年 5 月ごろからは、月に 100 回も行わないことが多くなってきているからである。

6. 結論

日本語版 Wikipedia では、管理者が編集を行う割合は年々減少している傾向だった。

参考文献

- [1] Kevin Kelly. The bottom is not enough, 2008. <http://kk.org/thetechnium/the-bottom-is-n/>. 堺屋七左衛門訳. ボトムアップだけでは不十分. <http://memo7.sblo.jp/article/16282081.html> (2016.1.25 閲覧) .

農業における Wiki を活用する知の構造化

プロジェクトマネジメントコース 矢吹研究室 1242034 氏名 小池 克人

1. 序論

平成 27 年 3 月 10 日に発表された農業情報の標準化に関する個別ガイドラインでは、農業の情報の相互運用性を確保するインターオペラビティーとポータビリティ確保の標準化が必要とされている [1].

しかし、異なる企業や団体の意思を一つにまとめる作業には困難が伴う。標準化したデータをやり取りするプロトコルやデータ形式は各企業の利害がぶつかるため、困難である [2]. 故に複数のシステムの間でマスターを統一しようとする共通語彙は、目的と実現する考え方のアーキテクチャーが異なる。マスターは構造も用語も異なるため標準化することができない。今回は共通的な方法論に統一することができないかを考える。

作物名称は、研究や行政などにより変わる。研究は研究目的に応じた分類のため、行政は行政の目的に応じた分類のため、などの用途により視点と用語が異なる。

今回の研究では、異なる作物名称を目的に応じた用語に変換（翻訳）するための、共通する仕組みがないかということに着目した。用語の変換をするために用語を上位下位で結び、関連用語を抽出することを考える。関連用語の抽出をしている事例を探し、参考にして研究する。そして、自然言語処理に Wikipedia 上位下位関係抽出の論文を参考にして、Mediawiki を利用する [3]. MediaWiki は、Wikipedia に使用されているフリーソフトで自分の Wiki を自分のサーバーに設置することができる。そのため MediaWiki を利用し、目的に応じて最適な語彙の翻訳を可能とする翻訳システムの開発を研究する。

2. 目的

本研究では、農業情報の用途により用語が異なる語彙を目的に応じた最適な語彙への翻訳ができるような仕組みを作ることが目的である。

3. 手法

本研究では、MediaWiki を用いて知識を登録する。登録した情報から、上位下位関係抽出ツールを用いて単語間の関連情報を抽出をする。抽出した情報を使って、用語の翻訳を試みる。

4. 結果

Mediawiki に様々な書き方をした解析することで成功するパターンを発見することに成功した。カテゴリに出荷統計作物名などを記載して、記事の見出しのレベルを上げて書くことで目標とする解析結果を表示できた。

5. 考察

成功パターンを発見できたため、結果に記載したように Wiki を編集すれば、Wiki を編集したことがない人でも容易に編集することができ、農業情報の標準化ができると考えられる。

6. 結論

本研究では、上位下位関係抽出ツールを用いて、関連情報を抽出することに成功した。どのように運用していくかが今後の課題である。

参考文献

- [1] 内閣官房情報通信技術 (IT) 総合戦略室. 農業情報の標準化に関する個別ガイドライン等について. https://www.kantei.go.jp/jp/singi/it2/senmon_bunka/nougyou_guideline/siryoul.pdf.
- [2] 井原敏宏. 第 2 段階に入る農業 IT、複数企業・組織が連携して標準化や協業を推進, 2015. <http://itpro.nikkeibp.co.jp/atcl/column/14/346926/032900208/?ST=system>.
- [3] 隅田飛鳥, 吉永直樹, 鳥澤健太郎. 自然言語処理による wikipedia からの上位・下位関係抽出. 自然言語処理, Vol. 16, No. 3, pp. 3–24, 2009.

オンラインショッピングサイト利用者による商品に対するレビューの動向調査

プロジェクトマネジメントコース 矢吹研究室 1242042 齋藤 勇也

1. 序論

インターネットを利用した電子商取引は 1994 年に米国のピザハットが行ったのが最初であるといわれている [1]. 1994 年以前, 商品の購入方法は商品の下に足を運ぶ必要があった. つまり商品を購入した人物は直接顔を合わせた相手にのみにしかレビューを語ることが出来ない状態である.

書籍によると全ての商取引における電子取引の割合が 2014 年時点で 3.7 % となり, 2008 年の 1.8 % 比べ 倍近く上昇し, 仮想空間でも商品の売買が行いやすい環境であることが分かる [2]. 仮想空間での商品の売買が可能となった 1994 年以降は, 電子商取引であるオンラインショッピングのレビューが重要視されている.

レビューが実装されている有名なオンラインショッピングサイトでは, 利用者は商品についてのレビューを記入することや, 商品に得点を付けることが可能である. 例えば, Amazon のレビューではおすすめ度と称して平均値しか表示していない. 商品とは無関係のレビューや商品の特徴を理解せず記述したレビューがあり, 平均値などの統計に加えるべきでないものが存在する.

そこで, レビューの表示方法が平均評価では商品の判断材料としての指標が少ないと判断し, 平均値よりも信頼できる方法を探す.

2. 目的

Amazon はレビューを平均値で表示している. 他の指標を加え判断材料を増やすことで, 平均値のみの表示よりも信頼できるレビューを作る.

3. 手法

各レビューがどれだけ信用できるかの指標として, 参考になったと答えた人物の比率, 購入者のレビューのみでの参考になったと答えた人物の比率を計測する.

4. 結果

平均との差異を図るために計 86 件のレビューデータを収集し, 散布図を作成した. 平均評価と重み付き評価の間で相関があり, 相関係数は 0.97 という数値が算出された. また, Amazon 内で購入した, していない人物を分類分けした. 9 件のレビューデータを収集し, 平均評価と購入した人物のみのレビューを使用した重み付き評価の散布図を作成し, 相関係数は 0.5 という数値が算出された.

5. 考察

重み付き評価値は相関係数が 0.97 であり, 購入者のみの重み付き評価値は相関係数 0.5 で購入者のみの相関係数のほうが相関が低いと言える.

6. 結論

レビューデータ全体から分析する方法に比べ, 相関が低い結果が出た購入者に絞ったレビューデータから参考になったと答えた比率を分析すれば平均値より正確な商品の評価がわかる可能性が高いと言える.

参考文献

- [1] 菅坂玉美, 横尾真, 寺野隆雄, 山口高平. e ビジネスの理論と応用. 東京電機大学, 第 1 版, 2003.
- [2] 経済産業省. 電子商取引実態調査. http://www.meti.go.jp/policy/it_policy/statistics/outlook/H25_summary.pdf (2014/11 時点).
- [3] 服部哲弥. Amazon ランキングの謎を解く. 化学同人, 第 1 版, 2011.
- [4] 山澤美由起, 吉村宏樹, 増市博. Amazon レビュー文の有用性判別実験. 情報処理学会研究報告自然言語処理 (NL), Vol. 2006, No. 53, pp. 15–20, may 2006.

クラウドファンディングにおける成功の判別分析

プロジェクトマネジメントコース 矢吹研究室 1242105 三浦泰介

1. 序論

クラウドファンディングが世界中で利用されている [1]. クラウドファンディングとは、プロジェクトの活動資金を、インターネットを利用し、不特定多数の支援者から募集する資金調達の手法である。

クラウドファンディングは一般的に資金提供者に対するリターンの形態によって、金銭的リターンのない「寄付型」と金銭的リターンのある「投資型」、権利や物品を購入することで支援する「購入型」に分けられる。

日本におけるクラウドファンディングは、2014年に金融商品取引法が改正されるまで [2] は、見返りを得ない寄付型、購入型に限られていた。購入型はリターンがあるためリターンを目当てに多くの出資者が集まる傾向があり、購入型は日本で一番市場が大きいクラウドファンディングの形態である。

2. 目的

プロジェクトにおいて資金集めは重要で、資金集めができなければプロジェクトは破綻してしまう。クラウドファンディングの成否に関わる要因を見いだすことでプロジェクトの資金集めでの成功率を上げることを目標とする。

3. 手法

クラウドファンディングの成功要因を以下の手順で分析する。

まず、日本の大手クラウドファンディングサービスである Makuake と READYFOR を定期的 (1 日 1 回) チェックし、そこで実施されているクラウドファンディングについての情報を収集する。収集する情報は、目標金額と支援コース数、支援最低金額、支援最高金額、資金提供者が得る物、動画の有無である。最終的に資金調達に成功したかどうかとも記録する。

次に、資金調達の成否の決定木を作成する。決定木の目的変数は資金調達の成否、説明変数は上述の収集情報 (目標金額等) である。

最後に、作成された決定木をもとに、クラウドファンディングの成功要因を考察する。

4. 結果

100 個のプロジェクトのデータを収集し、決定木分析を行った。使用した要因は「手法」で述べたとおりである。作成した決定木は本文に載せる。この決定木は、全体の 83% を再現している。

5. 考察

決定木から、支援最低金額が 3,744 円から 35,990 円で目標金額が 375,000 円以上のプロジェクトの成功率が最も高くなることがわかった。これは支援最低金額でリターンがもらえるプロジェクトが成功しやすいためだと思われる。

目標金額が 375,000 円以上の場合の成功率が高くなることは、この目標金額が今回調査したプロジェクトの平均目標金額 (160,6724 円) よりかなり低く、そのことが資金調達のしやすさに影響しているためだと思われる。最低金額と目標金額が決定木に頻出することからも、この 2 つの設定額が重要であることがわかる。

6. 結論

クラウドファンディングにおいて、プロジェクトの内容以外の成功要因を、決定木分析によって調査した。決定木の予測が実際の結果とよく合っていることから、そのような要因が存在することが示唆される。

参考文献

- [1] 西山恵太. クラウドファンディングを活用した新製品開発. NRI パブリックマネジメントレビュー, Vol. 140, pp. 8–13, 2015.
- [2] 松尾順介. わが国のクラウドファンディング規制の現状. 証研レポート, No. 1690, pp. 1–19, 2015.

集合知の成功事例としての株価変動についての調査

プロジェクトマネジメントコース 矢吹研究室 1242109 三宅琢己

1. 序論

1986年スペースシャトル・チャレンジャー号爆発事故が起きた。その直後、事故に関連していた4社の株価が急落した。そのうち3社の株価は持ち直したのだが、1社の株価だけさらに下がり、持ち直すことはなかった。その数か月後、その1社の部品が原因で爆発事故が起きたことを公表した[1]。この結果だけ見ると株式市場は原因企業を特定していたのではないかと考えられる。株式市場は賢く、原因企業を本当に特定していたのか、それとも偶然このようなことが起きたのか。

2. 目的

背景にあることを調査するためには、銘柄・期間を指定し、さらに株価データの変動を可視化できるようにする必要がある。そのようなことを可能にするツールの開発を本研究での目的とする。完成次第背景にある、株式市場は賢いのか、そうでもないのかを進めるものとする。

3. 手法

文献[2]を参考に株価の取得ツールを開発した。そして開発したツールを試し、背景の調査を進める。今回、2015年10月以降問題となっている、「横浜市マンション傾斜問題」に関連している企業の株価を研究の対象とする。「横浜市マンション傾斜問題」に関連し、判明した日は以下のとおりである。

- 三井不動産
2015年10月14日マンション傾斜公表する
- 旭化成
2015年10月14日杭打ちデータ改ざん認める
- 三井住友建設
2015年10月24日設計より短い杭を発注したこと認める
- 日立ハイテクノロジーズ
2015年10月27日旭化成に下請けを丸投げ疑惑がかかる

これらの日付をもとに株価変動を比較していく。

4. 結果

関連企業の株価変動結果

- 三井不動産は13日から下がり始め、14日は2パーセント下落
- 旭化成は10月14日まで変動はなく、15日の始値は14日の終値から11.4パーセント下落
- 三井住友建設は13日終値から14日始値は12パーセント下落し、14日始値と終値で30パーセント下落、その後持ち直さなかった
- 日立ハイテクノロジーズはトレンドを乱すことなく、特に変化なし

5. 考察

株価に変動があったのは日立ハイテクノロジーズ以外の3社である。

旭化成は14日に杭打ちデータ改ざんを認めたため株価が下がるのは当たり前であると考える。

だが三井住友建設は10月24日に短い杭を8流していたことを認めたのにもかかわらず、14日に大幅にトレンドが乱れ下落している。この結果を見る限り、株式市場は三井住友建設が主犯であると言っているように感じられる。

6. 結論

本研究の目的である株価取得とその可視化は成功した。背景の調査に関してはマンション傾斜の事例だけを見れば株式市場は原因をわかっていたと言えるが、事例が一つしかないため断言はできない。開発したツールを使い、事例を増やし研究していけば調査は進められる。

参考文献

- [1] ジェームズ・スロウィツキー. 「みんなの意見」は案外正しい. 角川文庫, 2009.
- [2] 佐々木拓郎. Rubyによるクローラー開発技術. SBクリエイティブ, 2014.

大学入試センター試験数学 I・A を用いた数式処理システムの性能評価

プロジェクトマネジメントコース 矢吹研究室 1242116 森谷慧士

1. 序論

東京大学の入試問題を全自動で解くプロジェクト（東ロボプロジェクト）が進められている [1]. その処理システムは、大学入試センター試験の模擬試験（5教科8科目）で950満点中511点（偏差値は57.8）を獲得、数学と世界史の偏差値は特に高く、数学 I・A は偏差値 64、数学 II・B は偏差値 65.8、世界史は偏差値 66.5 であったという [2]. 人工知能がこのような発達すると、その影響は数学教育にも及ぶだろう。今日の数学教育は、すべてを紙と鉛筆で行うことを前提に行われているが、その一部はコンピュータで置き換えることができるはずである。人間が行うこととコンピュータが行うことをうまく識別する能力の育成が求められるようになるだろう。

2. 目的

本研究では、センター試験の数学 I・A を題材にして、数学教育にコンピュータを導入することの可能性を調査する。その結果として、高校程度の数学能力を問う問題にコンピュータを活用して解く際に必要となる数学の知識とコンピュータの知識を明らかにすることを目指す。

3. 手法

本研究では、センター試験の数学 I・A の問題をコンピュータを用いて解き、問題を解くために必要な数学の知識とコンピュータの知識を確認する。2009 年から 2015 年までの 7 年分のセンター試験の数学 I・A の問題を解く。本研究では、数式処理システムの中で最も普及しているものの一つである Mathematica を採用する。本研究では、無料で利用できるクラウドサービス版、Wolfram Programming Lab^{*1}を用いる。問題を解いたら、その際に用いた Mathematica の組み込みシンボルの種類を数える。センター試験の数学 I・A の問題を解くのに使える

組み込みシンボルが出尽くせば、シンボルの種類の増加は止まるはずである。

4. 結果

2009 年から 2015 年までのセンター試験の数学 I・A の問題を解き、利用した Mathematica の組み込みシンボルの累積数を記録した。利用した組み込みシンボルは全部で 23 個となる。

5. 考察

組み込みシンボルの累積数の増加が図累積グラフのように 6 年で止まったため、センター試験の数学 I・A の問題を解くのに必要な Mathematica の機能は、「結果」に記載した 23 種類で十分であることがわかる。Mathematica のような数式処理システムには膨大な機能が備えられているが、数学 I・A のために必要なのはこのように比較的少数の機能があり、教育の現場に導入するのも容易だと思われる。

6. 結論

本研究では、数式処理システム Mathematica を用いてセンター試験の数学 I・A の問題を解いた。その際に利用した Mathematica の組み込みシンボルを集計することによって、センター試験の数学 I・A を解くのに必要な数式処理システムについての知識を明らかにすることができた。本研究のような事例を増やすことによって、コンピュータのサポートのもとで数学の問題を解くのに必要な数学の知識を明らかにすることが今後の課題であろう。

参考文献

- [1] 新井紀子. ロボットは東大に入れるか. イースト・プレス, 2014.
- [2] ITmedia ニュース. 人工知能「東ロボくん」、センター試験模試で「偏差値 57.8」 数学と世界史は偏差値 60 超え. <http://www.itmedia.co.jp/news/articles/1511/16/news061.html> (2015.08.10 閲覧) .

^{*1} <https://lab.open.wolframcloud.com/objects/wpl/GetStarted.nb>

SNS 経由で入手される情報のユーザ間差異の可視化

プロジェクトマネジメントコース 矢吹研究室 1242131 吉野聡志

1. 序論

世界的に人気のある SNS(Social Networking Service) のひとつとして Twitter が存在する。2015 年 9 月 30 日現在、月間アクティブユーザは 3 億 2000 万人である [1]。

Twitter の速報性や拡散性を活かし、様々な企業が Twitter を用いたマーケティングを行っている。その例のひとつに、江崎グリコ株式会社が販売する「ポッキー」がある。11 月 11 日を「ポッキー&ブリッツの日」とし、毎年 11 月 11 日になると「ポッキー」とたくさんつぶやくよう呼びかけ、「24 時間で最もツイートされたブランド」としてギネス世界記録樹立を狙う。その取り組みで実際にギネス世界記録に認定されるだけでなく、ポッキーの売上が大きく跳ね上がるという結果も生み出した [2]。

SNS の中でもアクティブユーザ数が非常に多く、利用スタイルも多数ある Twitter に対し、ユーザである人々が持っているニーズが何であるかが分かれば、他の SNS との差別化を図りやすくなる。これにより効率的なマーケティングの手法を Twitter 社や、Twitter 上に広告を打ち出す企業に提案できるのではないかと考えられる。

2. 目的

数名の Twitter タイムラインを取得し、各タイムラインで特徴となる単語をリストアップする。第一の目的は、その語群が各人の Twitter での関心事項や利用スタイルとどの点が一致するか、またはしないかを見つけることである。

また第二の目的は、リストアップした単語や Twitter の利用スタイルをユーザごとに比較し、ユーザひとりひとりがどれだけ違った情報を Twitter 上でやり取りしているかを見比べることである。

3. 手法

まず、Twitter アクティブユーザである 5 人の矢吹研究室 3 年生に協力を仰ぎ、各人が Twitter に求める情報や、Twitter の利用スタイルを聞く。

次に自分のものを含む 6 つのアカウントにおい

て、同じ期間のタイムラインを一日分取得する。

そして取得したタイムラインを形態素解析エンジン MeCab で解析し、頻出単語とその頻度 (TF) を算出し、同時に各単語の重み付け (IDF) を行う。

最後に TF と IDF をかけ合わせ (TFIDF)、各タイムラインの特徴となる単語 (特徴語) と特徴の度合いを導き出す。

4. 結果

どのユーザにおいても、Twitter への関心事項とタイムラインの特徴語の一致をある程度観測することができた。

また、直接的な一致がなくとも関心事項に関連する単語も現れていた。

リストアップした単語のユーザによる違いは大きなものとなった。

5. 考察

Twitter でユーザが他のユーザをフォローする基準のひとつに、自分が興味のある情報をそのユーザが発信するかどうか、があると考えられる。

6. 結論

本研究の手法を用いて Twitter への顕在的なニーズを読み取るのは十分に可能といえる。しかし、ユーザが他のユーザをフォローする基準は人それぞれであり、それによつてはニーズと関係のない、ノイズとなるデータがどれだけ入ってくるかが変わるため、潜在的なニーズを予測する際は他の有用な情報も同時に得る必要がある。

参考文献

- [1] Twitter, Inc. Twitter, inc. について | about. <https://about.twitter.com/ja/company> (参照 2016-01-24) .
- [2] Members Co.,Ltd. 【twitter 運用】twitter 運用代表事例 4 選から考える、企業の twitter 活用のトレンドとこれから | コラム | メンバーズ. <http://blog.members.co.jp/article/14475> (参照 2016-01-25) .

GitHub 上のソフトウェア開発のためのフロー推薦手法

プロジェクトマネジメントコース 矢吹研究室 1242132 若月 純

1. 序論

ソフトウェア開発では、複数のメンバが同時に開発を行うため、様々な問題が発生する。問題を解決するため、バージョン管理システムを用いる。バージョン管理システムとは、変更履歴を管理するシステムのことである [1]。

バージョン管理システムを提供するサービスに、GitHub がある。GitHub は、バージョン管理システムに加え、開発を補助する機能を提供するサービスである。

GitHub を使用する手順を開発フローと呼ぶ。開発フローは、13 種類あり、それぞれメリットとデメリットがある。しかし、選択する基準は定められていないため、状況にあった開発フローを選択するのは難しい。そのため、適切でない開発フローを選択し、開発に悪影響を与える危険がある。

そこで、本研究では、適切な開発フローを選択できるようにするための基準を求める。

2. 目的

GitHub を用いたソフトウェア開発プロジェクトの性質において、適切な開発フローを選択できるようにするための基準を求める。

3. 手法

初めに、GitHub 上のプロジェクトから、開発フローの採用に関わるとされる指標を調査する。

次に、採用されている開発フローを調査する。開発フローの正解データは、人手で作成する。

最後に、プロジェクトに関する上述の指標から、開発フローを決めるための決定木の作成を試みる。

4. 結果

決定木分析結果が、図 1 である。

図 1 より、行数、バイト数、1 日あたりの行数で、選択されている開発フローを割り出せる。

開発フローのわかっているプロジェクトを使って作成された開発フローの決定木が、開発フローが未知のプロジェクトの開発フローを予測できる

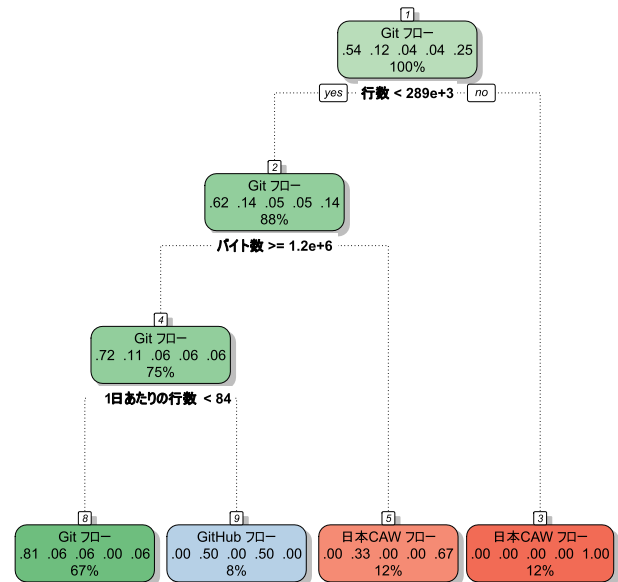


図 1 プロジェクトの性質により選択される開発フローの違い

かどうかを試したところ、精度は平均 38% (信頼区間は 33 ~ 42%), 再現率は平均 46% (信頼区間は 41 ~ 51%) だった。

5. 考察

図 1 の決定木は、1 日あたりの行数といった、時系列データで分岐している。ここから、他の時系列データを調査することで、決定木の精度と再現率をあげられるだろう。

6. 結論

本研究では、決定木を用いた、開発フローを推薦する手法を提案した。現状では、精度と再現率が高いとは言えないが、このような手法を発展させることによって、GitHub の経験が少ないチームでの開発でも、最適なフローを決定できるようになることが期待される。

参考文献

- [1] 池田尚史, 藤倉和明, 井上史彰. チーム開発実践入門〜共同作業を円滑に行うツール・メソッド. 技術評論社, 2014.