修 士 論 文

回帰分析によるソーシャルブックマーク数の予測モデルの構築

北陸先端科学技術大学院大学
情報科学研究科情報科学専攻

高松 征賢

2012 年 3 月

修 士 論 文

回帰分析によるソーシャルブックマーク数の予測モデルの構築

指導教官　東条敏 教授

審査委員主査　東条敏 教授
審査委員　島津明 教授
審査委員　白井清昭 准教授

北陸先端科学技術大学院大学
情報科学研究科情報科学専攻

1010038 高松 征賢

提出年月: 2012 年 2 月

概 要

本論文では、被リンク数や RSS 購読者数などの外的要因に頼らず、Web ページの内容や文章の構造などから Web ページの有用性を予測する手法を提案する。具体的には、特定の単語が出現したかどうかや本文の文字数、リンクの数などを特徴量とし、ソーシャルブックマーク数を正解とした回帰分析を行うことにより、ソーシャルブックマーク数の予測モデルを構築する。この手法によって、まだ人には知られていないが有用である Web ページ群を自動的に発見・ランキング付けを行い、ユーザに提供するシステムや、ユーザが Blog などのシステム上で文章を作成した場合に書かれた文章が人気がでるかどうかを判定するシステムの実装、情報検索エンジンのランキング精度の向上等が可能になると期待される。

目 次

第1章　はじめに

1.1　背景と目的

近年、インターネット上に存在する情報は膨大な量となり、その中から目的の情報を見つけ出す情報検索技術は非常に重要なものとなっている。Google をはじめとした既存の情報検索システムでは、基本的に、多くの人に知られている、すなわち被リンク数の多い Web ページが重要なページとみなされ、検索結果の上位にランキングされる [1]。しかし、このような被リンク数に基づく手法では、インターネット上に大量に存在する、有用ではあるがまだ人に知られていない Web ページを発見することができないという問題がある。

また、近年、Web 上の有用なページをコミュニティとして管理、発見するためのサービスとしてソーシャルブックマークサービスが注目されている。ソーシャルブックマークサービスとは、オンライン上で不特定多数のユーザがブックマークを管理・共有できるサービスである。ユーザはブックマークを付ける際にタグやコメントなどの情報を付加することができる。ソーシャルブックマークは人手によって厳選された有用な情報であり、ソーシャルブックマークが付けられた数を Web ページが有用であるかどうかの指標として扱うことができる。しかし、ソーシャルブックマークサービスでも被リンク数を用いた情報検索システムと同様の問題を抱えている。すなわち人に知られていないが有用である Web ページを発見することはできない。

そこで本研究では、被リンク数や RSS 購読者数などの外的要因に頼らず、Web ページの内容や文章の構造などから Web ページの有用性を予測する手法を提案する。具体的には、特定の単語が出現したかどうかや本文の文字数、リンクの数などを特徴量とし、ソーシャルブックマーク数を正解とした回帰分析を行うことにより、ソーシャルブックマーク数の予測モデルを構築する。この手法によって、まだ人には知られていないが有用である Web ページ群を自動的に発見・ランキング付けを行い、ユーザに提供するシステムや、ユーザが Blog などのシステム上で文章を作成した場合に書かれた文章が人気がでるかどうかを判定するシステムの実装、情報検索エンジンのランキング精度の向上等が可能になると期待される。

1.2　本論文の構成

本論文の構成は以下の通りである。第 2 章では、PageRank、ソーシャルブックマーク、機械学習、回帰分析、予測について関連研究を述べる。第 3 章では、提案手法とシステムの実装方法について述べる。第 4 章では、評価実験の手順と評価方法、その結果、考察について述べる。第 5 章では、まとめと今後の課題について述べる。

第2章　関連研究

本章では、関連研究として、PageRank、ソーシャルブックマーク、機械学習、回帰分析、予測について述べる。

2.1　PageRank

Google をはじめとした既存の情報検索システムでは、Web ページの人気度によって検索結果のランキング付けを行なってきた。Google が使用している検索結果のランキング付けのアルゴリズムとして PageRank[1] がある。PageRank は、Page と Brin らによって提案されたアルゴリズムであり、Web ページがどこをリンクして、どこからリンクされているかというリンク構造からなるグラフのみを使用して Web ページの人気度を測っている。PageRank は以下の式で計算される。

$$R(P_u) = c \sum_{P_v \in B_{P_u}} \frac{R(P_v)}{|P_v|} \tag{2.1}$$

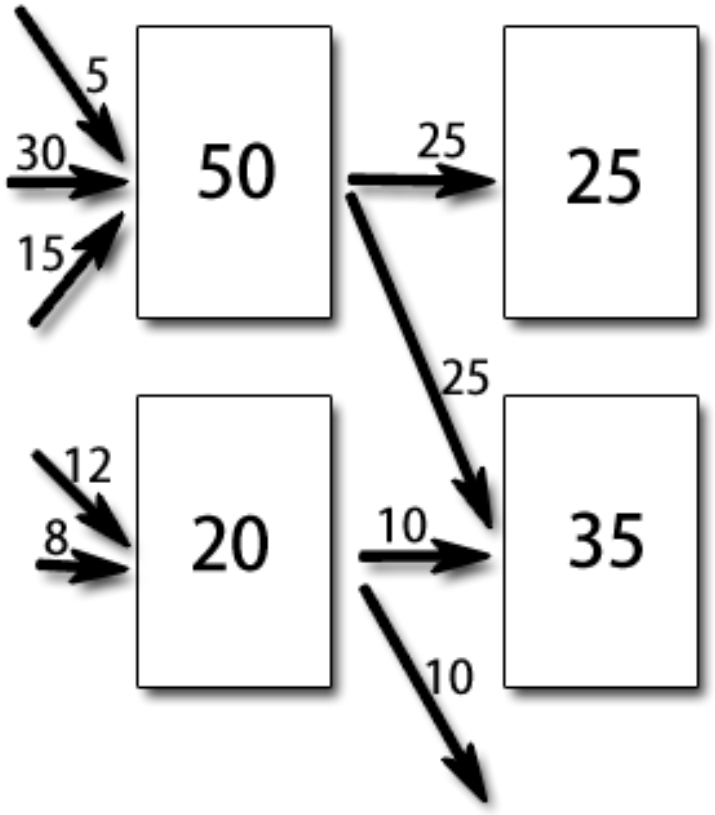
P はページ、R は PageRank、c は一般化を行うための定数、B_{P_u} はページ u をリンクしているページの集合、$|P_v|$ はページ v がリンクしている数の総数を表す時、$R(P_u)$ は P_u をリンクしているページの PageRank の総和となる。リンクしている数の総数で割っているので、ページ内にリンクが多く存在するページからのリンクは評価が小さくなっている。つまり、より多くの質の高いリンクを集めているページが人気がある Web ページとして評価付けられる。これを図で表すと図 2.1 のようになる。四角い図形が Web ページ、矢印がリンク、数字が PageRank を表している。

しかし、PageRank の手法では、リンクを使用しているため、人に知られていない Web ページや生成されたばかりの Web ページは評価することができない。

2.2　ソーシャルブックマーク

近年、Web 上の有用なページをコミュニティとして管理、発見するためのサービスとしてソーシャルブックマークサービスが注目されている。ソーシャルブックマークサービスとは、オンライン上で不特定多数のユーザがブックマークを管理・共有できるサービスである。

図 2.1: PageRank の例



丹波ら [2] は、ソーシャルブックマークを使った Web ページ推薦システムの研究を行なっている。丹波らは、ソーシャルブックマークと Folksonomy を利用してインターネット上の Web ページ全体を対象としたユーザに対する Web ページ推薦システムの構築手法を提案している。本研究の目的は、有用な Web ページを自動的に発見しランキング付けすることなので、ユーザに対する Web ページ推薦システムとは異なる。

毛受ら [3] は、ソーシャルブックマークサービスのユーザとブックマークに付与されたタグの時系列情報を利用して、新たに投稿された Web ページの注目度を予測している。ブックマークの増加の極大値をとる直前にページをブックマークしたユーザを高く評価し、予測モデルを構築している。ソーシャルブックマークを利用し、Web ページの注目度を予測するという点では、本研究と同じであるが、毛受らの研究ではブックマークサービスのユーザとタグ情報を利用しており、予測を行うために時間経過が必要である。一方、本研究では、そういった情報は使わずに Web ページの内容や文章の構造などから特徴量を抽出することで、時間を経過することなく Web ページの有用性を予測できるという点で異なる。

根本ら [4] は、Web ページ間のリンク構造に加えて、ソーシャルブックマークを付けたユーザ間の評価、タグ情報などを使用し、Web ページの玄人度を測ることで Web ページの評価を行っている。

Blei ら [5] は、supervised LDA の性能評価として、ソーシャルブックマークサービスの一つである Digg[15] のデータを使用して、Web ページの人気を予測している。

高橋ら [6][7] は、被ブックマーク数に時間変化を加え Web ページの質を測っている。

Golder ら [8] は、ページが最初に投稿されてから一両日中にソーシャルブックマーク数の増加のピークを迎え、増加のピークが始まる区間と終わる区間には加速度の大きな極大値と極小値が現れ、累積ブックマーク数はこの間に大きく伸びると指摘している。

2.3 機械学習、回帰分析、予測

Support Vector Machine (SVM) とは、教師あり学習を用いた線形二値分類器である [9]。また、カーネル法と組み合わせることで、非線形な分類も行うことができる。Support Vector Regression (SVR) とは、SVM を回帰分析が行えるように拡張したものである [10]。本研究では、回帰モデルとして SVR モデルを使用する。

Hong ら [11] は、twitter[16] のツイートの情報などを機械学習することによりそのツイートがリツイートされるかどうかを予測している。リツイートとは、ツイートを引用することで、ツイートがどれだけリツイートされたかの数がそのツイートの人気の度合いとして測ることができる。予測に用いる特徴量として、TF-IDF、LDA、PageRank、Degree distribution、局所的クラスタ係数、相互リンク、ツイートとリツイート元までの時間、前のツイートからの時間、リツイートされている間隔の平均、リツイートされるまでの平均時間、以前にリツイートされたかどうか、ユーザのリツイートされた数、ユーザのツイートした数が使われていた。データとしては、2009 年の 11 月から 12 月の間に収集し

た、約1億のツイートと250万ユーザのデータを使用している。結果として、一番精度が良かった特徴量の組み合わせとして、TF-IDF、LDA、Degree distribution、以前にリツイートされたかどうか、ユーザのリツイートされた数があげられ、適合率は99.3%、再現率は43.5%を出した。

Wuら[12]は、線形回帰とロジスティック回帰を用いてeBayの売れる商品とそうでない商品の予測を行なっている。予測に用いる特徴量として、カテゴリの人気度・コンバージョン率、商品の単語の人気度（eBayサイト内の検索エンジンでよく検索される単語)、その商品と競合する商品の数（タイトルの類似度で判別)、その類似商品が売れた比率、価格（類似アイテムとの相対価格)、オークションであるかどうか（オークションでない場合は即決価格が設定されている)、出品者の他の商品の量、返品対応、値段交渉できるかどうか、店頭在庫かどうか、出品者の評価された数、出品者のすべての評価におけるポジティブの数、出品者のレベル（eBay独自の評価によるもの）が使われていた。データとしては、eBay商品の内、2008年5月から2008年6月中に終了した720,076商品を使用した。ロジスティック回帰の評価には商品が売れたかどうかの適合率、再現率、F値を計測している。ロジスティック回帰の結果として、売れた商品の適合率は67%、再現率は65%、F値は66.3%となり、売れなかった商品の適合率は70.5%、71.7%、71.1%となった。線形回帰の評価には、システムによって選ばれた上位N個のセットの内何個が売れたかを計測している。カテゴリの評価としては、ビデオゲーム、健康美容用品、携帯電話、スポーツ用品、本等のカテゴリに対して比較的良い精度を出していた。また、特徴別の評価としては、類似商品が売れた比率、類似商品の数、出品者のすべての評価におけるポジティブの数、カテゴリの人気度等が比較的良い精度を出していた。

第3章　回帰分析によるソーシャルブックマーク数の予測モデル

本章では、回帰分析によるソーシャルブックマーク数の予測モデルの手法概要や、予測に用いる特徴量についての解説、本手法によるシステムの実装についての解説を行う。

3.1　手法概要

本研究では、被リンク数等の外的要因を使わずに Web ページに含まれる内容や文字数、文章の構造などを特徴量とし、ソーシャルブックマーク数を正解とした、SVR モデルで回帰分析を行い、個々の記事に付与される一週間後のソーシャルブックマーク数の予測を行ことで、Web ページの有用性を予測する手法を提案する。すべての Web ページと、特定の Blog サービスに書かれた Web ページを分類し、別々の回帰モデルを用いて実験を行う。

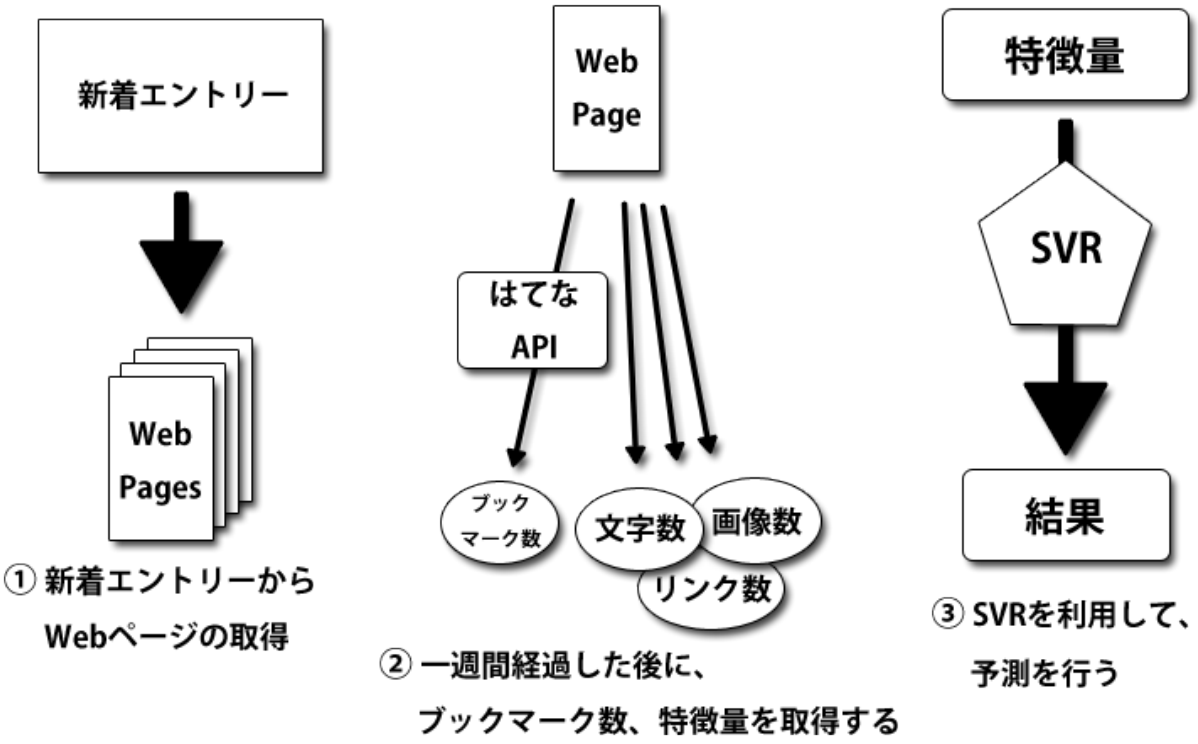
3.2　システムの実装

システムの流れとしては、はてなブックマークの新着エントリー [18] から Web ページ群を取得する。次に、ブックマークが十分に付けられる一週間後に、はてなブックマークエントリー情報取得 API[20] を使用してブックマーク情報を取得する。次に、Web ページの HTML から特徴量を抽出する。このとき特定の Blog サービスの記事であった場合には本文から特徴量も抽出する。最後に、回帰モデルとして、SVR を利用して、個々の記事に付与される一週間後のソーシャルブックマーク数の予測を行う。図 3.1 は本システムの処理の流れを図に表したものである。

3.2.1　Web ページの収集

はてなブックマークの新着エントリー [18]（閾値 3）に掲載された上位 10,000 件の Web ページを 2012 年 1 月 1 日から 2012 年 1 月 7 日までの間に一日一回クローリングを行い取得した。新着エントリーには、閾値以上のブックマーク数を獲得した Web ページのうち、一番初めに付けられたブックマークから閾値までのブックマークが付くまでの時間が早い

図 3.1: システムの流れ


新着エントリー
Web Pages
① 新着エントリーから
Webページの取得
Web Page
はてな API
ブック マーク数
文字数
画像数
リンク数
② 一週間経過した後に、
ブックマーク数、特徴量を取得する
特徴量
SVR
結果
③ SVRを利用して、
予測を行う

ページから順番に並んでいる。したがって、毎日異なった 10,000 件の Web ページを収集できる訳ではなく、複数日クローリングを行うと重複ページが存在してしまうので、最終的にはこの 7 日間で 20,505 件の Web ページを収集することができた。

3.2.2 本文の抽出

Blog などの Web ページでは、ユーザが主にコンテンツとして読む本文の部分と、リンクや広告が表示されたサイドバーやコメント欄など本文以外の部分で分けることができる。ユーザは本文の記事の内容を読むために Web ページに訪れているため、本文の特徴量を考慮するべきであるので、本文の部分のみから特徴量を抽出することで精度を向上させることができるかどうかを検証するために本文部分の抽出を行う。実験では、Web ページ全体から特徴量を抽出した場合と、本文のみから特徴量を抽出した場合とで、二通りの実験を行い比較、検証を行う。特定の Blog サービスとして、はてなダイアリー[1]、はてなブログ[2]、ライブドアブログ[3]、アメーバブログ[4]、FC2 ブログ[5]、Seesaa ブログ[6]、goo ブログ[7]、ココログ[8]、Yahoo!ブログ[9]、エキサイトブログ[10]のいずれかに含まれる Blog サービスを使用した Web ページの記事を使用する。それぞれの Blog サービスで本文を表すタグ内に含まれる文章を本文として抽出した。

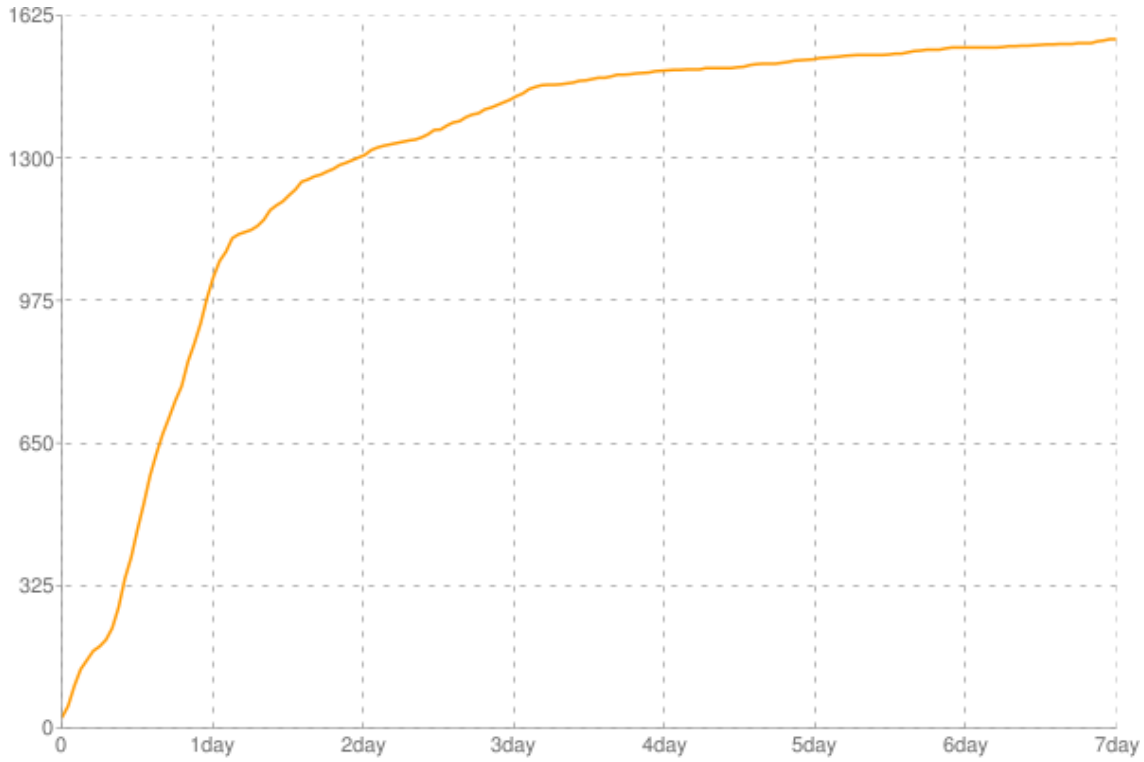
3.2.3 時間経過によるブックマーク数の推移

記事が投稿されてから、どのくらいの時間が経てばブックマーク数が十分に付き予測が行えるようになるかを調べるために、実際にいくつかの Web ページの一週間のブックマーク数の時間経過による推移を検証してみた。データとしては、ブックマーク数が 100

[1]はてなダイアリー
http://d.hatena.ne.jp/
[2]はてなブログ
http://hatenablog.com/
[3]ライブドアブログ
http://blog.livedoor.com/
[4]アメーバブログ
http://ameblo.jp/
[5]FC2 ブログ
http://blog.fc2.com/
[6]Seesaa ブログ
http://blog.seesaa.jp/
[7]goo ブログ
http://blog.goo.ne.jp/
[8]ココログ
http://www.cocolog-nifty.com/
[9]Yahoo!ブログ
http://blogs.yahoo.co.jp/
[10]エキサイトブログ
http://exblog.jp/

以上を獲得しているWebページを5つ抜き出し使用した。以下の図3.2[11]、図3.3[12]、図3.4[13]、図3.5[14]、図3.6[15]は、それぞれのサイトに、はじめてブックマークが付けられた時間から一週間後までの間に付けられたブックマーク数の推移を、横軸を時間、縦軸をブックマーク数としてとった折れ線グラフである。

図 3.2: Webページのブックマーク数の時間の経過による推移 その1



一両日中にWebページに付けられたブックマーク数の7,8割程度まで増加し、7日目にはほぼ増加しなくなることから、本研究ではWebページに最初にブックマークが付けられてから一週間後のブックマーク数であれば、十分な予測を行うことが可能であると見て、一週間後のブックマーク数を予測することとする。

[11]驚くほど違う→あなたの文章を最適化するたった4つのルール 読書猿Classic: between / beyond readers
http://readingmonkey.blog45.fc2.com/blog-entry-562.html
[12]リンク：論文に死んでも書いてはいけない言葉30 - 発声練習
http://d.hatena.ne.jp/next49/20120103/p2
[13]404 Blog Not Found:コードについて書く方がコードを書くより読まれる現実
http://blog.livedoor.jp/dankogai/archives/51767934.html
[14]大学の教員になりたい全ての人のために - 生駒日記
http://d.hatena.ne.jp/mamoruk/20120104/p1
[15]テキストファイルを処理するときのUnixコマンドまとめ - nokunoの日記
http://d.hatena.ne.jp/nokuno/20120121/1327139192

図 3.3: Web ページのブックマーク数の時間の経過による推移 その 2

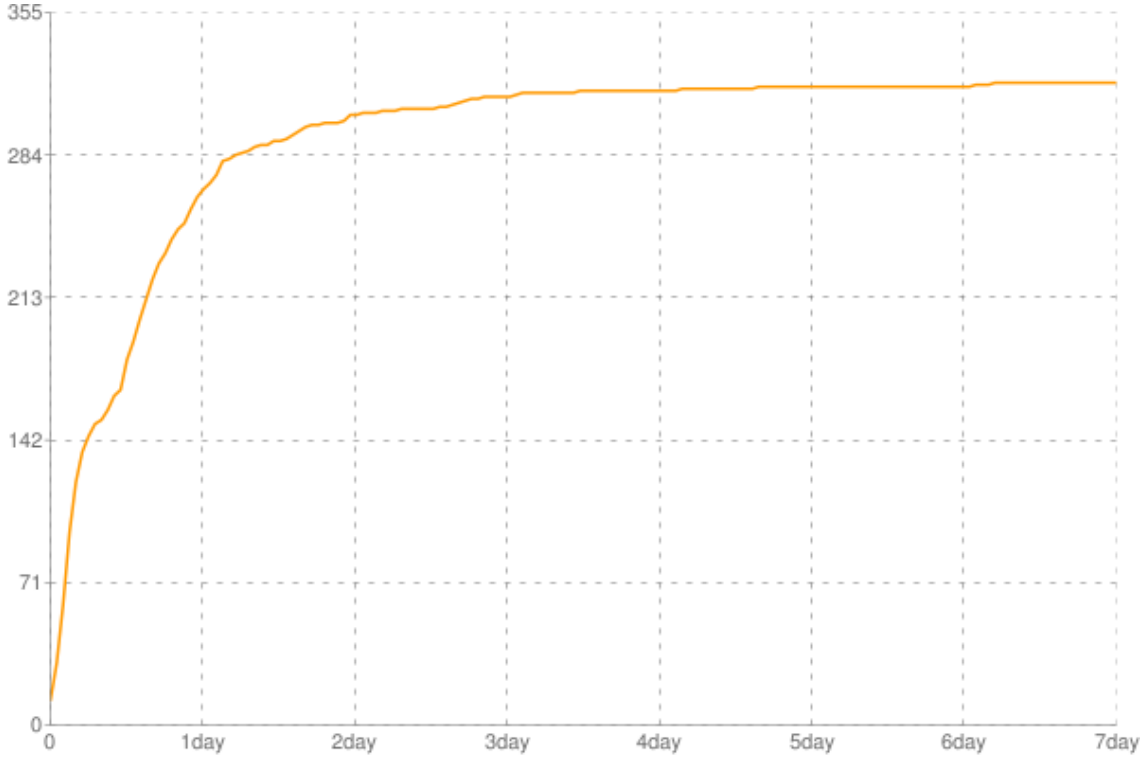


図 3.4: Web ページのブックマーク数の時間の経過による推移 その 3

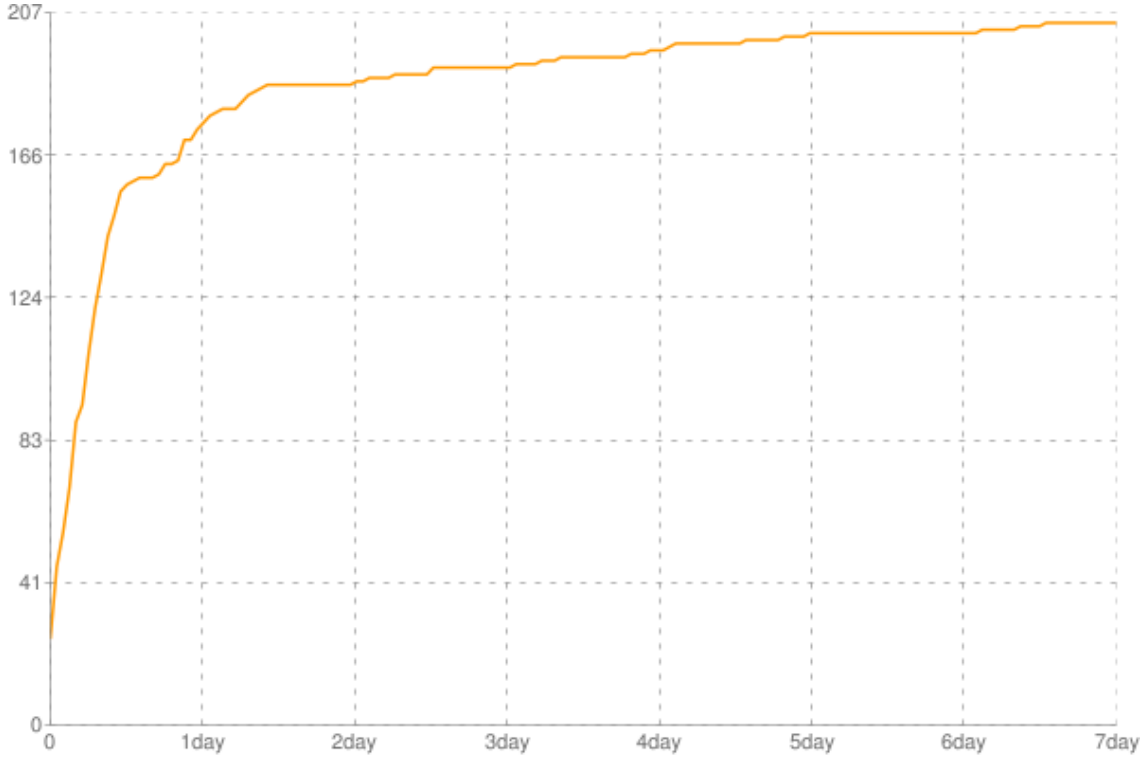


図 3.5: Web ページのブックマーク数の時間の経過による推移 その 4

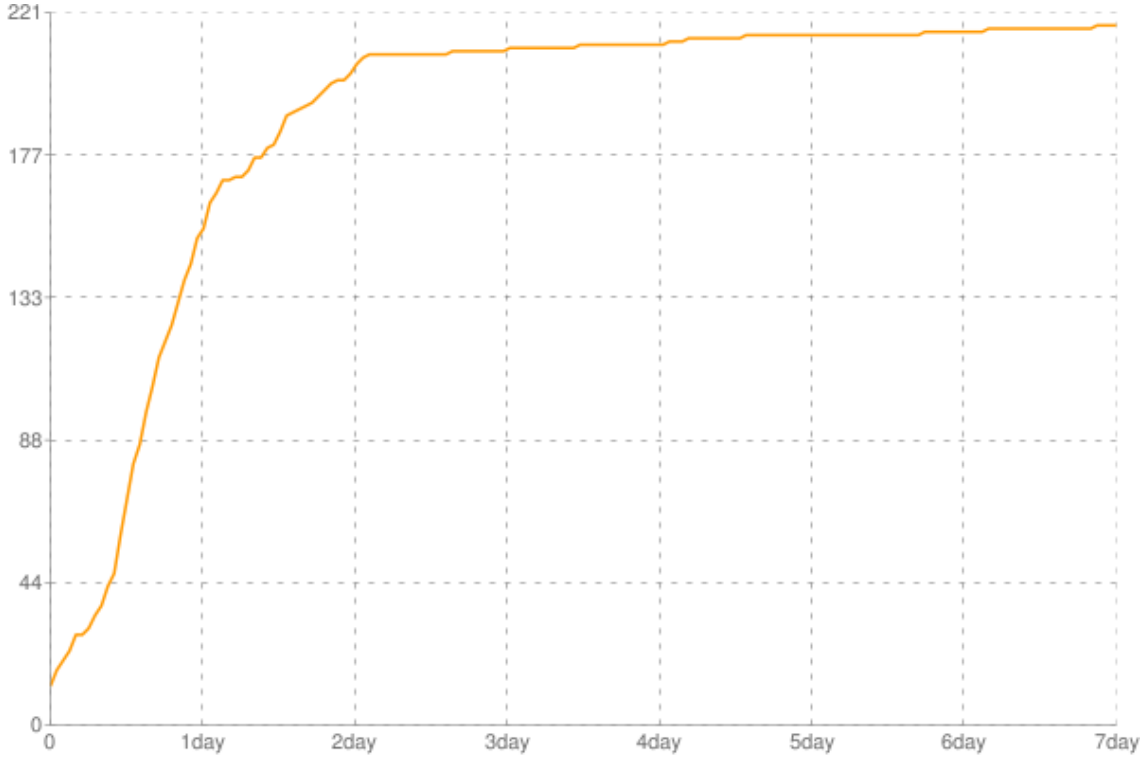


図 3.6: Web ページのブックマーク数の時間の経過による推移 その 5

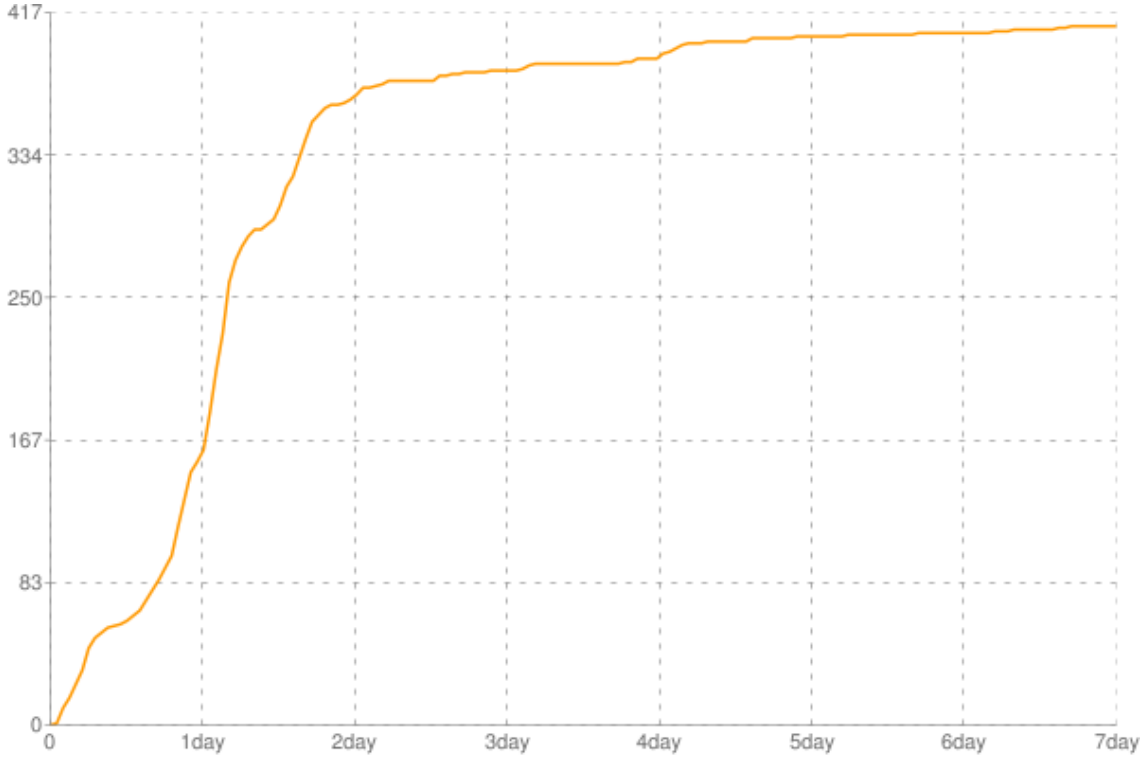


表 3.1: はてなブックマークエントリー情報取得 API JSON データ構造

名前	説明
title	タイトル
count	ブックマークしている合計ユーザ数
url	ブックマークされている Web ページの URL
entry_url	はてなブックマークのエントリーページの URL
screenshot	スクリーンショット画像の URL
eid	エントリー ID
bookmarks	ユーザがブックマークしたデータの配列。配列の構造は表 3.2。

表 3.2: bookmarks 配列の構造

名前	説明
user	ブックマークしたユーザの名前
tags	ブックマーク時に付けられたタグの配列
timestamp	ブックマークされた時刻
comment	ブックマーク時に付けられたコメント

3.2.4 ブックマーク情報の取得

ソーシャルブックマークのデータには、日本最大級のソーシャルブックマークサービスである、はてなブックマークのデータを使用する。ブックマーク情報の取得には、はてなブックマークエントリー情報取得 API[20] を使用する。

以下の URL のリクエストパラメータにブックマーク情報を取得したい Web ページの URL を送信することにより、JSON 形式でデータを取得することができる。ただし、プライベートユーザの情報は取得することができない。

“ http://b.hatena.ne.jp/entry/jsonlite/?url= ”

JSON データの構造は表 3.1 の通りである。3.2.3 項で記述したように、Web ページに最初にブックマークが付けられてから一週間後のブックマーク数であれば、十分な予測を行うことが可能であると見て、Web ページデータ取得から一週間後に、はてなブックマークエントリー情報取得 API を使用し、ブックマーク情報を取得する。

3.2.5 MeCab による形態素解析

タイトルやテキストを単語単位に分割し特定の単語が出現したかどうかの特徴量として加えるために、MeCab[14] を用いて形態素解析を行った。MeCab とは、オープンソースソフトウェアの形態素解析エンジンであり、パラメータの推定に Conditional Random Fields (CRF) を用いている。バージョンは MeCab 0.98 を使用、辞書には IPA 辞書を使用。

3.3 特徴量

予測に用いる特徴量について解説する。

3.3.1 予測に用いる特徴量

- 単語出現

  Web ページ中に含まれる単語が出現したかどうかの特徴。Web ページの内容によって Web ページが有用であるかどうかが左右されると考えられることから、単語が出現したかどうかを特徴とする。3.2.5 項で解説した方法で MeCab[14] で Web ページに含まれる文章に形態素解析を行い単語単位に分割し、名詞、動詞、 形容詞、連体詞、感動詞のいずれかの品詞にあてはまる単語を特徴として使用する。出現した単語には一意な ID を付け 1 次元とする。

- タイトルの単語出現

  タイトルに含まれる単語が出現したかどうかの特徴。すべてのユーザが本文をすべて読みブックマークを付けているわけではなく、見出しやタイトルのみを見て判断し、ブックマークを付けるユーザもいると考えられる。したがって、タイトルの内容を判断するために、この特徴を使用する。単語出現の特徴と同じ手順で、タイトルに出現した単語を特徴量として加える。

- 文字数

  少ない情報を掲載している Web ページよりも、よりたくさんの情報を掲載している Web ページの方が有用であるだろうと考えられることから、文字数を特徴とする。Web ページに含まれる文字数を数えて特徴量として加える。

- 画像

  HTML の画像の挿入に使用する ⟨img⟩ タグが出現した回数を画像の数とする。画像の数と文字数に対する画像の数の割合を特徴量として加える。

- リンク

  多数のリンクが付けられている Web ページはスパムである可能性が考えられることから、リンクの数とリンクの文字数それらの割合を特徴とする。HTML のリンクを表す ⟨a⟩ タグが出現した回数をリンクの数とする。また、⟨a⟩ タグで囲まれた文字の数をリンクの文字数とする。リンクの数とリンクの文字数と文字数に対するリンクの数の割合、文字数に対するリンクの文字数の割合を特徴量として加える。

- 改行

  多数の改行が行われている Web ページは有用ではなさそうという考えから改行の数と割合を特徴とする。HTML の改行を表す ⟨br⟩ タグが出現した回数を改行の数とする。改行の数と文字数に対する改行の数の割合を特徴量として加える。

- 文字種類

  Web ページには最適な文字種類の比率で書かれた文章の方が読みやすくユーザにとって有用なページであるのはでないか（例えば、ひらがなが文章の大半を占めるテキストは読みにくくユーザにとって有用なページではないのではないか。）、と考えられることから文字種類の比率を特徴量として採用する。ひらがな、カタカナ、漢字の割合を特徴量として加える。

- 流行語

  はてなブックマークのサイト内の人気エントリー [19] に掲載された Web ページから流行語を抽出し特徴量として加える。人気エントリーのタイトルを抽出し、MeCab でタイトルを単語単位に分割し、単語の出現頻度を数える。人気エントリーのタイトル内と同じ単語がタイトルに出現していた場合には、その単語の出現回数を特徴量として使用する。n は単語の ID、x は日付、x_n は日付 x 中の n の出現回数とした場合、以下の式により計算する。

$$R_x = \sum_{n=0}^{n} x_0 + x_1 + \cdots + x_n \tag{3.1}$$

例えば、ある日の人気エントリーに“インターネット”という単語が3回、“まとめ”という単語が5回出現していた場合では、“便利なインターネットサービスまとめ”というタイトルには、“インターネット”と“まとめ”が一回ずつ含まれているので、流行語 R_x の特徴量は 8 となる。しかし、以上の方法では、“-”や“|”のようなタイトルと Blog 名の区切りに使われている記号であったり、“ブログ”や“日記”という一般的な単語が高い出現頻度になってしまう。一般的に頻出する単語ではなく、数日間だけ話題となり頻出した単語を抽出し流行語としたいので、流行語を抽出したい日付から 1 週間前までに出現した単語の出現頻度の平均をとり、その日付の単語を平均で割ることで流行語の抽出を行う。特徴量の値は以下の式で計算される。

$$R_x = \frac{\text{式 } 3.1}{\frac{\sum_{n=0}^{n}(x-1)_0+(x-2)_0+\cdots+(x-6)_n}{7}} \tag{3.2}$$

以上の 2 つの方法で流行語を抽出した特徴を特徴量として使用する。

- 新着エントリー

  はてなブックマークには、新着エントリー [18] というページが存在し、新たにブックマークされた Web ページが表示される。しかし、新着エントリーは新たにブックマークが付けられたすべての Web ページを上から順番に表示している訳ではなく、ブックマーク数の閾値が設けられており、ある Web ページに一番はじめに付けられたブックマークから閾値までの時間が早い順にソートされている。つまり、早く連続でブックマークがつけられた Web ページが上位に表示される。閾値は 3、5、なしと設けられているが、リクエストパラメータを変更することで任意の値に変更することができる。ユーザは新着エントリーから Web ページを探しブックマークすることも多く、上位に掲載された Web ページが多数のブックマークを獲得する場合があるので、ブックマーク数を 3 つと 5 つ獲得するまでの時間を特徴量として採用した。86,400[16]から、はじめにブックマークが付けられてから 3 ブックマークされるまでの時間と、5 ブックマークされるまでの時間をそれぞれ引いたものを特徴量として加える。マイナスになる場合は、値を 0 とする。この特徴は、Web ページの記事の内

[16]86,400 は 1 日を秒数で表した数字。60（秒）× 60（分）× 24（時間）= 86,400

容等のみから有用なページを探すという本研究の目的とそれるが、新着エントリーに掲載された場合にどういった効果があるのかを検証するため特徴量として採用した。

- ドメイン

  同一ドメインかどうかを特徴量として加える。例として、“http://seiken.tk/abc.html”と“http://seiken.tk/xyz.html”は、ドメインが同じなので、同一ドメインとみなす。ただし、“ http://blog.seiken.tk/ ”のようにサブドメインが異なる場合は別のドメインとみなす。また、Blog サービスのようにディレクトリ名にユーザの ID 等を入れユーザの Blog として識別しているような Blog サービスでは、ユーザ ID となるディレクトリまでを 1 つのドメインとしてみなす。ドメインに一意な ID を割り当ていき、ID1 つにつき 1 次元とする。この特徴は、Web ページの記事の内容等のみから有用なページを探すという本研究の目的とそれるが、元々人気のある Web ページからの記事はブックマークを集めるのではないかという考えから、同一ドメインからの投稿によりどのくらいブックマークの付き方に影響があるのかを調べるため特徴量として採用した。

3.3.2 特徴量の正規化

特徴によって値の大きさに違いができてしまい回帰分析がうまく行かないことがある。例えば、“文字数”の特徴は、1,000 を超える値が普通であるが、“画像の数”の特徴は 100 以内の値であることが多く、これらの値をそのまま特徴量として扱った場合、値の大きな特徴量にばかり評価をしてしまいうまくいかない。したがって、他の特徴量と大きさの差を無くすために特徴量の正規化を行った。ただし、文字列に対する画像の数やリンクの数等の割合が値として特徴量となっている特徴は、正規化を行わずにそのまま特徴量として扱う。一つ目の正規化の方法として、同じ特徴の中の一番大きい値ですべての値を割る方法である。特徴量を f、特徴量の最大値を MAX_f とした場合、式は以下のようになる。

$$\text{特徴量}_f = \frac{f}{MAX_f} \tag{3.3}$$

しかし、この方法の場合では、1 つの値が極端に大きい場合は、他の値が小さいところで推移してしまう。以上の問題を解決するための二つ目の正規化の方法として、偏差値を算出する方法である。平均を AVG_f、標準偏差を STD_f とした場合、式は 3.4 のようになる。0.5 を足すことで、平均が 0.5 となるようにしている。

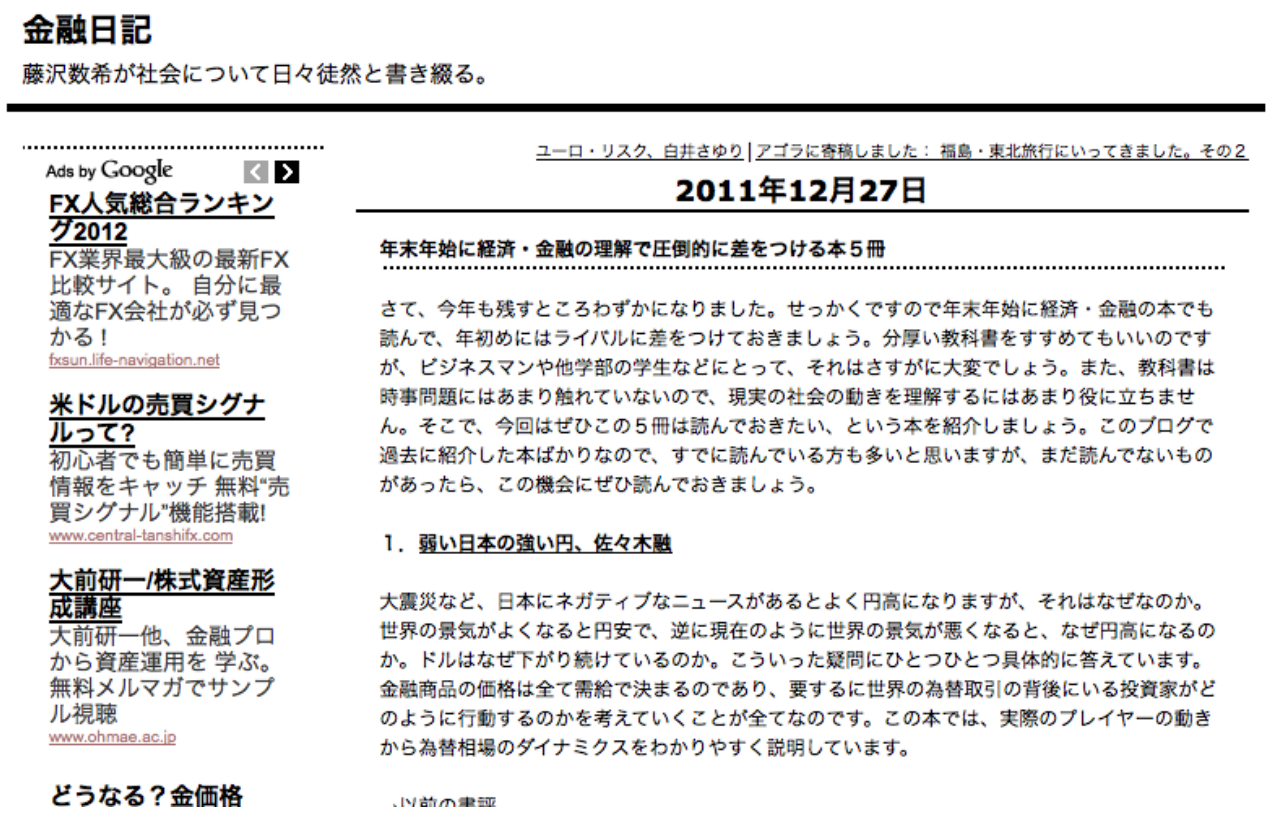
$$\text{特徴量}_f = \frac{\frac{f - AVG_f}{5}}{STD_f} + 0.5 \tag{3.4}$$

評価実験では、特徴量を最大値で割り正規化（式（3.3）を利用）した場合と偏差値を出し正規化 (式（3.4）を利用）した場合で二通りの実験を行い、どちらがより良い精度が出るかを検証する。

3.3.3 特徴量の抽出

本セクションでは、実際にある Web ページ[17]を例にとって特徴量の抽出を行う。図 3.7 が例にだした Web ページのスクリーンショット画像である。

図 3.7: Web ページのスクリーンショット



式（3.3）および式（3.4）による正規化で計算される特徴量を、それぞれ表 3.4、表 3.5 に示す。

表 3.3: 特徴量

特徴名	特徴量
ブックマーク数	689
blog 番号	784
文字数	13506
画像	22
リンク	37
リンク文字数	2545
改行	101
ひらがな	0.4422
カタカナ	0.0957
漢字	0.3056
3 ブックマーク	326
5 ブックマーク	1675
最初にブックマークされた時刻	2011-12-27 01:20:02

表 3.4: 最大値で正規化（式（3.3））した場合の特徴量

特徴名	特徴量
文字数	0.0101
画像	0.0058
リンク	0.0028
リンク文字数	0.0046
改行	0.0069
流行語 3.1	0.1083
流行語 3.2	0.1057
ひらがな	0.4422
カタカナ	0.0957
漢字	0.3056
3 ブックマーク	0.9962
5 ブックマーク	0.9806
文字数と画像数の比率	0.0016
文字数とリンク数の比率	0.0027
文字数とリンクの文字数の比率	0.188
文字数と改行の数の比率	0.0075

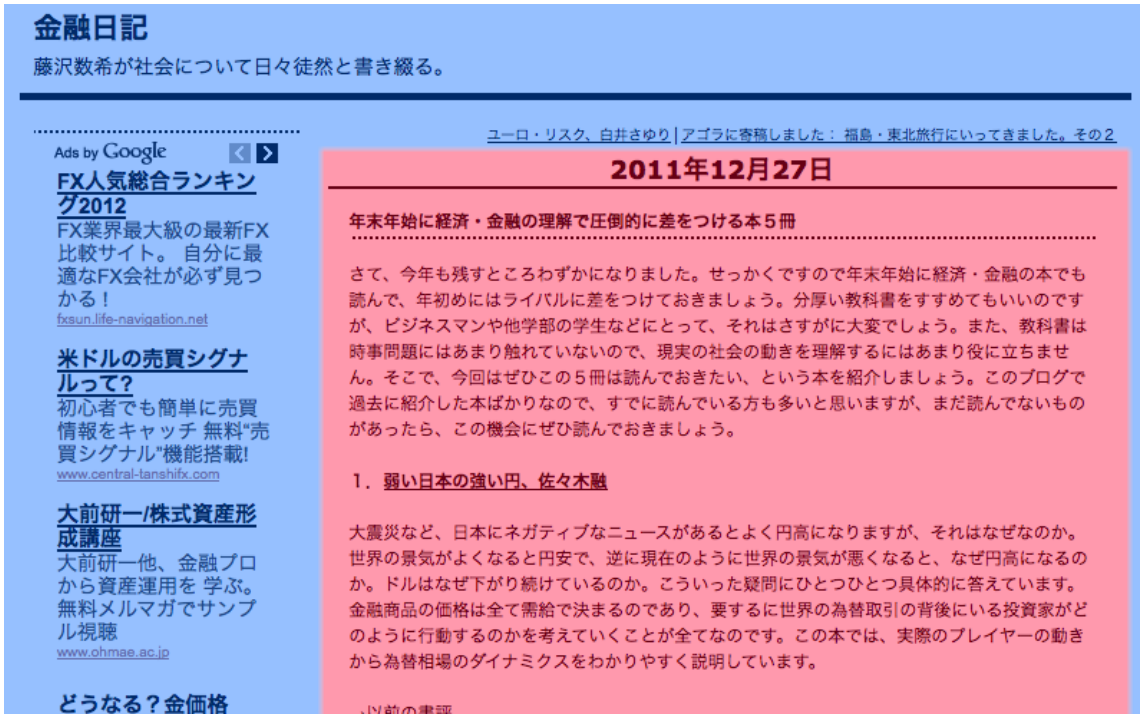
表 3.5: 偏差値で正規化（式（3.4））した場合の特徴量

特徴名	特徴量
文字数	0.4453
画像	0.374
リンク	0.4284
リンク文字数	0.3838
改行	0.4641
流行語（3.1）	0.4209
流行語（3.2）	0.6421
ひらがな	0.4422
カタカナ	0.0957
漢字	0.3056
3 ブックマーク	0.9962
5 ブックマーク	0.9806
文字数と画像数の比率	0.0016
文字数とリンク数の比率	0.0027
文字数とリンクの文字数の比率	0.188
文字数と改行の数の比率	0.0075

表 3.4 に加えて、タイトルと Web ページ中の単語が出現したかどうかの特徴量を加える。

特定の Blog（例の場合では、livedoor Blog であるので本文抽出を行う）の場合は本文のみを抜き出し、その中から特徴量を抽出する。図 3.8 は、赤い網掛けの部分が本文であるのでその部分の HTML を抽出する、青い網掛けの部分がそれ以外の部分である。

図 3.8: 本文とそれ以外の部分



本文のみを抜き出し、特徴量を抽出し、3.4 で解説した方法で正規化を行った結果が表 3.6 である。

以下が抽出された特徴量を SVR で機械学習するために成形された値である。

> 8.9542 1:0.0101 2:0.0058 3:0.0028 4:0.0044 5:0.0069 6:0.1083 7:0.1057 8:0.4422 9:0.0957 10:0.3056 11:0.9962 12:0.9806 13:0.0016 14:0.0027 15:0.0016 16:0.0075 800:1 4085:1 4137:1 4481:1 4569:1 4947:1 4962:1 4965:1 5147:1 5733:1 6461:1 6462:1 6463:1 300003:1 300004:1 300014:1 300020:1 300025:1 300026:1 300030:1 300038:1 300039:1 300040:1 300044:1 300052:1 300059:1 300072:1 300073:1 300083:1 300085:1 300092:1 300096:1 300098:1 〜Web ページ中の単語が出現したかどうかの特徴量が続く。

[17]金融日記:年末年始に経済・金融の理解で圧倒的に差をつける本５冊
http://blog.livedoor.jp/kazu_fujisawa/archives/51878633.html

表 3.6: 本文のみを抜き出し、偏差値で正規化（式（3.4））した場合の特徴量

特徴名	特徴量
文字数	0.4174
画像	0.4582
リンク	0.4529
リンク文字数	0.4259
改行	0.0069
流行語（3.1）	0.4209
流行語（3.2）	0.6421
ひらがな	0.5817
カタカナ	0.0843
漢字	0.2881
3 ブックマーク	0.9962
5 ブックマーク	0.9806
文字数と画像数の比率	0.0015
文字数とリンク数の比率	0.0024
文字数とリンクの文字数の比率	0.0015
文字数と改行の数の比率	0.0091

1 次元目が文字数、2 次元目が画像の数、3 次元目がリンクの数、4 次元目がリンクの文字数、5 次元目が改行の数、6 次元目が流行語（3.1）、7 次元目が流行語（3.2）、8 次元目がひらがなの比率、9 次元目がカタカナの比率、10 次元目が漢字の比率、11 次元目が 3 ブックマーク獲得までの時間、12 次元目が 5 ブックマーク獲得までの時間、13 次元目が文字数と画像の数の比率、14 次元目が文字数とリンクの数の比率、15 次元目が文字数とリンクの文字数の比率、16 次元目が文字数と改行の数の比率、17〜4,000 次元がドメインの ID、4,001〜300,000 次元がタイトルの単語出現、300,001〜次元が Web ページ全体の単語出現となっている。

第4章　評価実験

4.1　目的

実験の目的は、どの特徴が有用なWebページとなるのに有効であるのかを検証することである。また、記事の内容を単語出現の特徴量として採用しているが、タイトルとWebページ中全体の単語を対象にした場合では、精度に変化はあるのかを検証する。Blogなどの Webページでは、ユーザが主にコンテンツとして読む本文の部分と、リンクや広告が表示されたサイドバーやコメント欄など本文以外の部分で分けることができる。ユーザは本文の記事の内容を読みに Webページに訪れているので、本文の部分のみから特徴量を抽出することで精度を向上させることができるかどうか検証する。元々人気のあるWebページの記事がブックマークの増加に関係があるのかどうかを検証する。新着エントリーに掲載された場合にどういった効果があるのかを検証する。

4.2　実験

3.2.1 項で説明した方法で収集したデータを使用する。データの数は、総 Webページ数 20,505、総文字数 482,080,900、総ブックマーク数 362,233 となった。

- 実験 1

  実験では、ブックマークの値は、対数をとり値を縮めてある。評価方法として、真の値と予測して出した値を引いて絶対値をとり足しあわせ平均をとっているが、ブックマーク数の生の値の差の絶対値をとってしまうと、ブックマーク数が大きい場合の誤差を過大評価してしまうため、値の対数をとることで値を縮めている。実験 1 では、対数の底の値を 2、ネイピア数（e)、10 と変更した場合にどの値が一番良い精度がでるのかを検証する。

- 実験 2

  特徴量の正規化 3.3.2 で、最大値で値を割る式（3.3）による方法と値の偏差値を算出する式（3.4）による正規化の二通りを提案したが、実験 2 では、どちらの正規化の方法がより良い精度がでるのかを検証する。

- 実験 3

  実験 3 では、トレーニングデータの値を変更することで精度にどれくらい影響を与えるかを検証する。また、トレーニングデータとして、どのくらいの量を用意すればよいのかを調べる。

- 実験 4

  実験 4 では、個々の特徴量の有効性を調べるために、特徴量を種類ごとに削除して比較実験を行う。

- 実験 5

  実験 5 では、Web ページの記事の内容等のみから有用なページを探すという本研究の目的とそれるが、元々人気のある Web ページの記事がブックマークの増加に効果があるかどうかを調べるために、特徴量にドメインの ID を加える。また、新着エントリーに掲載された場合にどういった効果があるのかを検証するために、はじめにブックマークが付けられてから 3 ブックマークされるまでの時間と、5 ブックマークされるまでの時間の特徴量を加える。

- 実験 6

  実験 6 では、特定の Blog サービスで書かれた記事ページのみを抜き出し、そのページ全体を対象として特徴量の抽出を行なった場合と、本文部分のみを対象として特徴量の抽出を行なった場合とで比較を行い、精度の違いを検証する。

4.2.1 実験手順

テストデータとトレーニングデータには、Web ページにはじめにブックマークが付けられた時間が新しい Web ページから降順に並べ、上からテストデータとして 5,000 件、トレーニングデータとして 10,000 件の Web ページを使用した。つまり、新しい記事をテストデータとして使っており、実際に本システムを運用する場合には、古いデータから新しいデータの予測を行わなければならないため、現実問題に近い形で実験を行なった。学習器には、Joachims により提供されている SVM ソフトウェアである SVM^{light}[13] を用いて回帰モデルを構築し、実験を行なった。

- 実験 3

  トレーニングデータの数を 50、100、500、1,000、2,500、5,000、10,000 と増やしていく。

- 実験 4

  特徴量として、Web ページ中の単語出現、タイトルの単語出現、文字数、画像数、リンク数、リンクの文字数、改行数、文字種類、流行語、文字数と画像数の比率、文字数とリンク数の比率、文字数とリンクの文字数の比率、文字数と改行数の比率を使用する。そこから、特徴量を種類ごとに削除して比較実験を行い、どの特徴量が一番有効かを調べる。

- 実験 5

  特徴量に加えて、ドメインの ID、はじめにブックマークが付けられてから 3 ブックマークされるまでの時間と、5 ブックマークされるまでの時間を特徴量として加えた時の精度の変化を調べる。

- 実験 6

  収集したデータから特定の Blog サービス 3.2.2 に書かれた記事ページのみを抽出したところ、4,402 件の Web ページが抜き出された。20,505 件の Web ページの内、約 21.47%にあたる 4,402 件の Web ページが特定の Blog サービスに書かれた記事ページであったことがわかる。4,402 件の Web ページ全体の総文字数が 145,077,851 となり、本文のみの総文字数は 41,535,207 であったので、全体の約 28.63%が本文の量ということになる。初めにブックマークが付けられた日時が新しいものをテストデータとして 2,000 件を使用し、残りの 2,402 件をトレーニングデータとして使用する。

4.2.2 評価方法

予測誤差 4.1 の平均による評価を行う。ブックマークの真の値に 1 を足して[1]対数をとったものから、システムが予測した値に 1 を足して対数をとったものを引いて、絶対値をとり、その平均をとったものを予測誤差の平均として評価基準に使う。予測誤差の平均が小さいほど精度が良いということとなる。

$$\text{予測誤差} = \sum_{n=0}^{n} |log_x(1 + \text{真の値}_n) - log_x(1 + \text{予測した値}_n)| \tag{4.1}$$

また、もうひとつの評価方法として、ページに一定以上のブックマーク数が付くかどうかを予測する精度を評価する。評価基準としては、以下の式で定義される、適合率、再現率、F 値を用いる。

[1]ブックマーク数が 0 の Web ページも存在するので 1 を足し対数をとることができるようにする

$$Precision = \frac{\text{正解数}}{\text{システムが出した答え}} \tag{4.2}$$

$$Rcall = \frac{\text{正解数}}{\text{正解の総数}} \tag{4.3}$$

$$F = 2 \times \frac{(Precision \times Recall)}{Precision + Recall} \tag{4.4}$$

4.3 結果・考察

- 実験 1

  実験 1 では、対数の底の値を 2、ネイピア数（e)、10 と変化させ実験を行なった。結果としては、対数の底が 2 の場合の予測誤差の平均が 0.9607、ネイピア数（e）の場合が 0.6692、10 の場合が 0.2958 となった。これらの値を対数から実数に戻すと、2 の場合が 1.9462、ネイピア数（e）の場合が 1.953、10 の場合が 1.976 となり、対数の底が 2 の場合が一番予測誤差が小さかったので、これからの実験では対数の底の値として 2 を採用することとする。

- 実験 2

  実験 2 では、特徴量の正規化 3.3.2 の方法として、最大値で値を割る式（3.3）による方法と値の偏差値を算出する式（3.4）による二通りの正規化のうちどちらがより精度がでるのかを検証した。結果としては、最大値で値を割る方法の予測誤差の平均が 0.9611 となり、値の偏差値を算出する方法が 0.9607 となった。若干ではあるが値の偏差値を算出する方法の方が精度が良かったのでこれからの実験では特徴量の正規化の方法として値の偏差値を算出する方法を採用する。

- 実験 3

  トレーニングデータの数を増減させた場合の予測誤差の平均と、ブックマーク数が 50 を超えたかどうかを予測した値に 1.5 を足したものの適合率、再現率、F 値を調べたところ、表 4.1 の結果となった。トレーニングデータの数を増やしていったところ、予測誤差の平均、適合率、再現率、F 値すべての値が改善されていったが、トレーニングデータの数が 2,500 を超えると改善する割合は少なくなっていったので、トレーニングデータの数としては最低 2,500 以上用意するのが良いとわかった。

表 4.1: 実験 3 の予測結果

トレーニングデータの数	予測誤差の平均	適合率	再現率	F 値
100	1.0385	5.0%	0.8%	1.6%
500	1.0047	26.0%	7.2%	11.3%
1,000	0.9825	36.1%	8.4%	13.6%
2,500	0.9664	32.8%	15.2%	20.8%
5,000	0.9637	32.8%	16.0%	21.5%
10,000	0.9603	34.1%	17.9%	23.5%

- 実験 4

  すべての特徴量を使用した場合、予測誤差の平均が 0.9603 となり、ブックマーク数が 50 を超えたかどうかを予測した値に 1.5 を足したものの適合率は 34.1%、再現率は 17.9%、F 値は 23.5%となった。すべての特徴量を使用した場合から、特徴量を種類ごとに削除して比較実験を行なったところ表 4.2 の結果が得られた。特徴量を種類ごとに削除しているので、特徴量の削除を行い実験を行なった結果出された予測誤差の平均、適合率、再現率、F 値が悪くなっているほど、有効な特徴量であると言える。結果として、ページ中の単語出現が一番良い結果となり、次にタイトルの単語出現が良い結果となった。流行語は僅かでは他の特徴量よりは良い結果となった。それ以外の特徴量は Web ページの有用性には、ほぼ関係が無いということがわかった。

- 実験 5

  実験 4 で使用した特徴量に、ドメイン番号、3 ブックマークが付くまでの時間、5 ブックマークが付くまでの時間の特徴量を加えた実験を行なった。結果としては、表 4.3 のようになった。3 ブックマークが付くまでの時間、5 ブックマークが付くまでの時間、つまり新着エントリーにどれだけ早く掲載されたかの特徴量を加えた場合、予測誤差の平均、再現率、F 値が実験 4 で使用した特徴量よりも良い結果となった。ドメイン番号もブックマークが付くまでの時間の特徴量ほどではないが、実験 4 で使用した特徴量の予測誤差 0.9603 よりも良くなっているので有効であるということがわかる。以上のことから、3、5 ブックマーク付くまでの時間が早いほど新着エントリー上位に掲載されるので、ブックマークが付きやすくなるということがわかった。

- 実験 6

  ページ全体のテキストを対象とした場合の予測誤差の平均は 1.3408 となり、本文のみを対象とした場合の予測誤差の平均は 1.2854 となった。ブックマーク

表 4.2: 実験 4 の予測結果

削除した特徴量	予測誤差の平均	適合率	再現率	F 値
ページ中の単語出現	0.9821	29.0%	10.3%	15.2%
タイトルの単語出現	0.9638	33.3%	17.1%	22.6%
文字数	0.9603	33.8%	17.9%	23.4%
画像数	0.9604	33.3%	17.5%	22.9%
リンク数	0.9603	34.1%	17.9%	23.5%
リンクの文字数	0.9603	34.1%	17.9%	23.5%
改行数	0.9603	33.8%	17.9%	23.4%
流行語	0.9606	34.1%	17.9%	23.5%
文字種類	0.9603	33.8%	17.9%	23.4%
画像数の比率	0.9603	33.8%	17.9%	23.4%
リンク数の比率	0.9603	34.1%	17.9%	23.5%
リンクの文字数の比率	0.9603	34.1%	17.9%	23.5%
改行数の比率	0.9603	33.8%	17.9%	23.4%

表 4.3: 実験 5 の予測結果

加えた特徴量	予測誤差の平均	適合率	再現率	F 値
3、5 ブックマークまでの時間	0.5742	31.9%	53.2%	39.9%
ドメイン番号	0.9599	34.1%	17.9%	23.5%
両方	0.5740	32.0%	53.2%	40.0%

表 4.4: 実験 6 の予測結果

評価対象	予測誤差の平均	適合率	再現率	F 値
ページ全体	1.3408	2.2%	1.7%	1.9%
本文のみ	1.2854	9.1%	3.4%	5.0%

数が 82 を超えたかどうかを予測した値に 1.5 の値を足しあわせた結果の適合率、再現率、F 値は表 4.4 のようになった。図 5.19 と図 5.20 はそれぞれの適合率、再現率、F 値を折れ線グラフで表したものである。ページ全体よりも本文のみを対象とした方が予測誤差が小さく、適合率、再現率、F 値の精度がすべて良かったので、ユーザは本文の内容を見てページが有用であるかどうかを決めているので本文のみを抜き出し特徴量を抽出した方がより良い精度が出るという仮説は正しかったと言える。

第5章　終わりに

5.1　まとめ

本研究では、被リンク数などの外的要因に頼らずに、Web ページに内容や文章の構造などから Web ページの有用性を予測する手法を提案し、実験によりその有効性を示した。具体的には、Web ページ中に特定の単語が出現するかどうか、またタイトル中に特定の単語が出現するかどうか、ソーシャルブックマークの人気ランキングである、はてなブックマークの人気エントリーに掲載されたページのタイトルから流行語を抽出し、流行語がタイトル中にどれだけ含まれるか、特定の Blog サービスの記事ページに対しては、本文のみを抽出し別の回帰モデルで学習するとより良い結果が得られる。すべての特徴量を使用し SVR モデルで回帰分析を行なったところ、対数領域での予測誤差の平均は 0.9603 となり、ブックマーク数が 50 を超えたかどうかを予測した値に 1.5 を足したものの適合率は 34.1%、再現率は 17.9%、F 値は 23.5%となった。システムの予測誤差の平均、閾値以上の予測が行えたかどうかの適合率、再現率、F 値を比較することにより、今回提案した手法が有効であることがわかった。

5.2　今後の課題

今後の課題としては、今回使用した特徴量以外にも Web ページの有用性を測るような特徴量はあると考えられるので、他の特徴量も検証することである。例えば、背景の色や、文字の色、リンクの見やすさ、文字の大きさ、行間、ユーザが迷わないようなサイトマップの設計等ユーザが読みやすいようにユーザインタフェースを設計してあるかどうか。文章の簡潔さ（冗長ではない）、内容の専門性、内容の一貫性、他の Web ページには存在しないような内容であるかどうか等、より文章の内容や構造に踏み込んだ特徴量が Web ページに有用性に有効であるのではないかと考えられる。今回は、有用な Web ページを自動的に発見しユーザに提供するシステムの開発を行なったが、今後はユーザが Blog などのシステム上で文章を作成した場合に書かれた文章が人気がでるかどうかを判定するシステムの実装であったり、情報検索エンジンのランキング精度の向上等を行なっていきたい。

謝辞

本研究を進めるにあたり鶴岡慶雅先生、東條敏先生には様々なご指導をして頂き大変お世話になりました。また、研究室の先輩、同級生には研究に関する助言をして頂きました。お世話になった皆様に心から御礼を申し上げます。

参考文献

[1] L. Page, S. Brin, R. Motwani and T. Winograd. The pagerank citation ranking: Bringing order to the Web. Stanford Digital Library, Technical report, 1998.

[2] 丹波, 土肥, 本位田. Folksonomy マイニングに基づく Web ページ推薦システム. 情報処理学会論文誌, 47(5), pp. 1382-1392, 2006.

[3] 毛受, 吉川. ブックマークの時系列情報を利用したソーシャルブックマークにおける注目度予測. 電子情報通信学会 第 19 回データ工学ワークショップ, 2008.

[4] 根本, 後藤, 金井. ソーシャルブックマークにおけるタグ付けを利用した Web ページ評価手法の検討. 情報処理学会研究報告, pp. 55-60, 2009.

[5] Blei, David M. and Mcauliffe, Jon D. Supervised topic models. Advances in Neural Information Processing Systems 21, pp. 121-128, 2007.

[6] 高橋, 北川. ソーシャルブックマークにおけるブックマークの活性度を考慮した Web ページのランキング. データ工学と情報マネジメントに関するフォーラム, A4-1, 2009.

[7] 高橋, 渡邊, 北川. ソーシャルブックマークにおけるトピック分析と活性度推定に基づく Web ページのランキング. データ工学と情報マネジメントに関するフォーラム, D2-5, 2010.

[8] S. A. Golder and B. A. Huberman. The structure of collaborative tagging system. Information Dynamics Lab, HP Labs 2005.

[9] T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Proceedings of ECML-98, 1998.

[10] AJ. Smola, B. Scholkopf A Tutorial on Support Vector Regression. NeuroCOLT Technical Report TR Royal Holloway College, London, UK, 1998.

[11] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. Proceedings of WWW 2011 - Poster, pp. 57-58, 2011.

[12] X. Wu, A. Bolivar. Predicting the Conversion Provability for Items on C2C Ecommerce Sites. Proceedings of CIKM '09, pp. 1377-1386, 2009.

[13] T. Joachims. Making Large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, pp. 169-184, MIT-Press, 1999.

[14] 工藤拓. MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://mecab.sourceforge.net/.

[15] Digg. http://digg.com/.

[16] Twitter. http://twitter.com/.

[17] はてなブックマーク. http://b.hatena.ne.jp/.

[18] 新着エントリー. http://b.hatena.ne.jp/entrylist.

[19] 人気エントリー. http://b.hatena.ne.jp/hotentry.

[20] はてなブックマークエントリー情報取得 API. http://developer.hatena.ne.jp/ja/documents/bookmark/apis/getinfo.

付録

図 5.1: 開発したシステムの予測結果を表示するインターフェース

図 5.2: すべての特徴量を使用した場合の出力結果上位 30 件

ブックマーク数	予測した値	タイトル
7	94	今日、娘が天国に：キニ速　　気になる速報
208	93	痛いニュース(ﾉ∀`)：東洋大・柏原がAKB篠田の前で「AKB興味ない」発言　スタジオ凍りつく - ライブドアブログ
51	89	今インターネットでステマが大流行!!　Googleトレンドでステマが1位へ:【2ch】ニュー速VIPブログ(`・ω・´)
118	60	pixivで絵の仕事依頼をする側、される側が知っておきたいこと ｜ お絵かき速報！萌え絵上達法
68	60	痛いニュース(ﾉ∀`)：「ストラディバリウス」などバイオリン名器の音色、現代モノと大差なし？ - ライブドアブログ
92	55	痛いニュース(ﾉ∀`)：「無料のはずなのに10万円請求された」…無料うたった課金ゲームに消費者庁動く - ライブドアブログ
42	54	痛いニュース(ﾉ∀`)：韓国人歌手が東京で従軍慰安婦のポスター1500枚を掲示…関西地方でも予定 - ライブドアブログ
48	54	カオスちゃんねる：「ピラミッドは現代技術でも再現するのは至難の業」　←は？石積み上げればいいだけだろ
609	52	絵描く工程を晒していく - まめ速
63	47	マツコ・デラックス「韓国人でてけ！日本が嫌だったら！出てけ」:ハムスター速報
3	46	サービスの価格と価値を再考する: Kanamori Marketing Office
316	45	１万円くらいで始められる趣味:ハムスター速報
91	43	脳にまつわるsungeeeeeeeeってなる話:哲学ニュースnwk
100	41	漫画やアニメで登場人物の私服が糞ださいと「うわあ…」ってなるよね:【2ch】ニュー速VIPブログ(`・ω・´)
66	41	橋下市長「高齢者でなく、現役・将来世代に税を投入して強化。その活力によって高齢者を支える」:ハムスター速報
32	39	【2ch】ニュー速クオリティ:【これは全力で流行らせるべき】 無毛文化いよいよ定着か？　パイパンにする日本人女性が急増！
48	39	JR町田駅前にコスプレ集団出現ｗｗｗｗｗｗｗｗｗｗｗ:ハムスター速報
121	39	13兆……未納給食費の何年分だろう？ - Apes! Not Monkeys!　本館
35	38	暇人＼(^o^)／速報：【ブラウザ戦争】Google Chrome（27%）がIEとFirefox（25%）のシェアを奪い続けている　恐ろしい子 - ライブドアブログ
89	37	女子中学生バトル漫画始まる:ハムスター速報
47	37	箱根駅伝2区・東農大 木下淵哉の部屋がヤバイｗｗｗｗｗｗｗｗｗ:ハムスター速報
54	36	痛いニュース(ﾉ∀`)：【TBS】 新春ロードショー・映画「おくりびと」視聴率3.4% - ライブドアブログ
26	35	【2ch】ニュー速クオリティ:国内企業、クロマグロを5649万円で落札し4500万円以上の赤字を出す。「これはお金の問題ではない」
65	34	真価が問われる2012年のスマートフォン - ケータイ Watch
310	33	電子書籍の「探しにくさ」について « マガジン航[kɔ:]
296	33	「okama」は絵が上手い。独特のデザインセンスと色彩感覚が魅力的:【2ch】ニュー速VIPブログ(`・ω・´)
40	32	暇人＼(^o^)／速報：【宗教】オーディオ用LANケーブル発売 これで綺麗な音楽データがダウンロードできる！ - ライブドアブログ
18	32	不連続への挑戦 - Nothing ventured, nothing gained.
37	32	アクアリウムやってる奴こい:ハムスター速報
106	31	カオスちゃんねる：【閲覧注意】異世界に行く方法

図 5.3: すべての特徴量を使用した場合の出力結果下位 30 件

ブックマーク数	予測した値	タイトル
10	3	震災で中止、でもどうしても年内にやりたかった。プリキュアと歌手30人が3D映像を再現コンサートレポ(エキサイトレビュー) - エキサイトニュース
3	3	秩父行ってきたんで現地で撮った写真うｐする ： しまぱん
5	3	「3DS Lite」今年の春に登場か！？「厚み」や「駆動時間」を改良した軽量バージョン：オレ的ゲーム速報@刃
3	3	【画像】ＡＫＢ４８・倉持明日香ちゃんの野球モノマネンゴｗｗｗｗｗｗｗ WASHI NOTE -ワシノート-
4	3	初音ミクproject miraiのCMが初公開！あまりのかわいさに死者続出:萌えオタニュース速報
3	3	山本彩 ひこにゃん城	NMB48オフィシャルブログpowered by Ameba
3	3	サッカーミックスジュース：五輪代表にOA枠は要らない、"チームスポーツ"の意味と難しさを考えろ
5	3	２ちゃん的韓国ニュース：「日本人よ、罪なき人を何千万人殺したのか」～[寄稿]日帝のつらい記憶、その時間の意味
3	3	政経ch - 【政治】野田首相「日本の平和と繁栄と国民の幸福、特に被災地の復興をお祈り申し上げた」･･･4閣僚と伊勢神宮参拝
3	4	エロゲ情報とりあえずまとめ:フロントウイング「グリザイアの迷宮」体験版に無修正データ混入→差し替え
5	4	ソニー・他、ニュー速で懲りずにおまんちんに引っかかる：有題無題ゲハにゅ
3	4	幸か不幸か家族計画：嫁は料理が美味い。それは認める。
6	4	ぶりそく！：今の小学生が使ってる彫刻刀かっこよすぎワロタｗｗｗｗｗ
5	4	2期『これはゾンビですか? OF THE DEAD』先行映像でOP&ED公開！秋に「めっちゃフェスティボー（仮）」を開催:萌えオタニュース速報
3	4	【映画】立川シネマシティ　秋山澪お誕生日記念『映画けいおん！ライブスタイル上映』1月15日開催:萌えオタニュース速報
4	4	石原都知事「芥川賞候補がバカみたいな作品ばかりで読むのがしんどい。選考委員だから読むけど」：はちま起稿
5	4	中国人ロリ美少女のヌード写真その１：エロ画像・美少女ロリ専門ブログ
11	4	東亜速報＋
6	4	元代表GKの楢崎正剛が板野のCMを馬鹿にしてる件	AKBωタメ AKB48情報サイト
4	4	ナンピンしますたm9(｀・ω・´) 民主「住宅の固定資産税値上げするわ」　購入厨死亡ｗｗｗｗｗｗｗｗ
20	4	第十五回ゲスト：杉作J太郎（前編） - 吉田 豪の雑談天国（ニューエストモデル風） - コラム	Rooftop
5	4	帰省したら女子高生の妹に３０代の彼氏がいたんだが: フライドチキンは空をとぶ -フラソラ-
11	4	ぶりそく！：『スマイルプリキュア！』キャスト＆主題歌決定！ 福圓美里、金元寿子、井上麻里奈、田野アサミ、西村ちなみ、大谷育江
9	4	政経ch - 車買う金が無いお前らに朗報！20万円で買える車ｷﾀｰ･ﾟ･--.o..o.--･ﾟ･(∀)･--.o.. .o.--･ﾟ･--!!!!
3	4	なにこれ凄すぎる。イリュージョンすぎる平面交差が話題。奇跡の交差点。:1000mg
4	4	U-1速報：ヒュンダイ韓国車の詐欺的販売方法が鬼女によって広められつつあるようです
7	4	いすゞのトラックの歌、無駄にいい歌すぎだろ
5	4	【速報】松屋の「豚めし」販売終了：オレ的ゲーム速報@刃
13	4	上空から見たコミケが凄い！コミックマーケット81待機列の微速度撮影映像
8	4	seiyu fan：【イロエリ！】声優 喜多村英梨 色別キャラ一覧表が素晴らしい

図 5.4: 特徴量の例

```
3 1:0.4177 2:0.4415 3:0.4078 4:0.4066 5:0.4074 6:0.2947 7:0.4022 8:0.234 9:0.18 10:0.2528 13:0 14:0 15:0 16:0
24532:1 300014:1 300018:1 300038:1 300039:1 300044:1 300052:1 300057:1 300060:1 300146:1 300235:1 300256:1 300258:1
300355:1 300375:1 300427:1 300440:1 300450:1 300467:1 300469:1 300475:1 300501:1 300553:1 300645:1 300736:1 300800:1
300845:1 300863:1 300875:1 300933:1 301125:1 301170:1 301186:1 301365:1 301456:1 301499:1 301532:1 301576:1 301592:1
301616:1 301812:1 301860:1 301914:1 302039:1 302163:1 302351:1 302521:1 302600:1 302738:1 302810:1 302902:1 303084:1
303117:1 303210:1 303377:1 303459:1 303472:1 303473:1 304461:1 304556:1 305260:1 305363:1 305825:1 305951:1 305998:1
306547:1 306697:1 307055:1 308266:1 309852:1 310147:1 310175:1 310826:1 311369:1 314175:1 314545:1 314980:1 315373:1
316130:1 320050:1 320356:1 320938:1 320939:1 320940:1 320941:1 320942:1 320943:1 320944:1 320945:1 320946:1
2 1:0.4302 2:0.4415 3:0.4135 4:0.4148 5:0.4216 6:0.5786 7:0.5139 8:0.4585 9:0.0755 10:0.3428 13:0 14:0.0002 15:0
16:0.0054 24965:1 300014:1 300026:1 300038:1 300039:1 300044:1 300052:1 300059:1 300060:1 300092:1 300112:1 300113:1
300121:1 300143:1 300170:1 300183:1 300187:1 300204:1 300242:1 300256:1 300269:1 300271:1 300274:1 300277:1 300320:1
300337:1 300338:1 300346:1 300355:1 300365:1 300370:1 300461:1 300477:1 300480:1 300487:1 300498:1 300528:1 300536:1
300561:1 300592:1 300604:1 300614:1 300664:1 300665:1 300720:1 300785:1 300845:1 300863:1 300867:1 300872:1 300874:1
300891:1 300913:1 300914:1 300922:1 300943:1 300979:1 301008:1 301038:1 301043:1 301046:1 301065:1 301072:1 301074:1
301106:1 301114:1 301192:1 301194:1 301219:1 301288:1 301293:1 301295:1 301317:1 301421:1 301428:1 301456:1 301486:1
301538:1 301576:1 301584:1 301608:1 301739:1 301748:1 301920:1 301921:1 301933:1 301940:1 302010:1 302020:1 302037:1
302039:1 302192:1 302224:1 302268:1 302286:1 302320:1 302339:1 302377:1 302411:1 302637:1 302671:1 302689:1 302975:1
303163:1 303407:1 303422:1 303438:1 303623:1 303735:1 303859:1 303861:1 303980:1 304062:1 304121:1 304209:1 304437:1
304513:1 304536:1 304678:1 304723:1 304796:1 304898:1 304932:1 304936:1 305121:1 305289:1 305430:1 305501:1 305534:1
305628:1 305746:1 305822:1 306101:1 306159:1 306211:1 306323:1 306404:1 306653:1 306831:1 306945:1 307137:1 307161:1
307162:1 307299:1 307463:1 307464:1 307674:1 307800:1 307969:1 307993:1 308432:1 308929:1 309026:1 309564:1 309565:1
309606:1 309797:1 309880:1 310003:1 310288:1 310388:1 311093:1 311337:1 311369:1 311415:1 311552:1 311928:1 312326:1
312421:1 312998:1 313150:1 313621:1 315044:1 317018:1 317712:1 318509:1 319606:1 319835:1 319936:1 320052:1 320440:1
320656:1 320707:1 320965:1 320980:1 320985:1 320991:1 322449:1 322459:1 327706:1 327849:1 328612:1 331508:1 332134:1
338039:1 338670:1 339367:1 339368:1 339369:1 339370:1 339371:1 339372:1 339373:1 339374:1 339375:1 339376:1 339377:1
339378:1 339379:1 339380:1 339381:1 339382:1 339383:1 339384:1 339385:1 339386:1 339387:1 339388:1 339389:1 339390:1
339391:1 339392:1 339393:1
2 1:0.629 2:0.4849 3:0.4867 4:0.589 5:0.6207 6:0.5628 7:1.0676 8:0.4179 9:0.1374 10:0.2583 13:0.0007 14:0.0007
15:0.0007 16:0.0183 25710:1 300003:1 300004:1 300005:1 300014:1 300025:1 300026:1 300038:1 300039:1 300040:1
300044:1 300052:1 300053:1 300057:1 300083:1 300113:1 300121:1 300124:1 300132:1 300150:1 300160:1 300165:1 300168:1
300171:1 300174:1 300183:1 300198:1 300204:1 300230:1 300235:1 300242:1 300256:1 300257:1 300269:1 300293:1 300311:1
```

図 5.5: 実験 3：トレーニングデータの数を 100 とした場合の精度

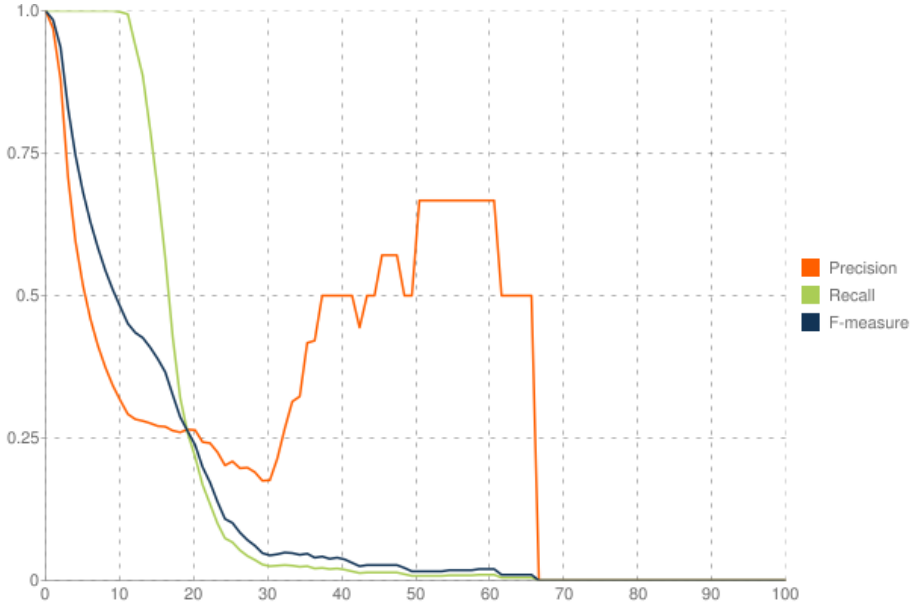


図 5.6: 実験 3：トレーニングデータの数を 500 とした場合の精度

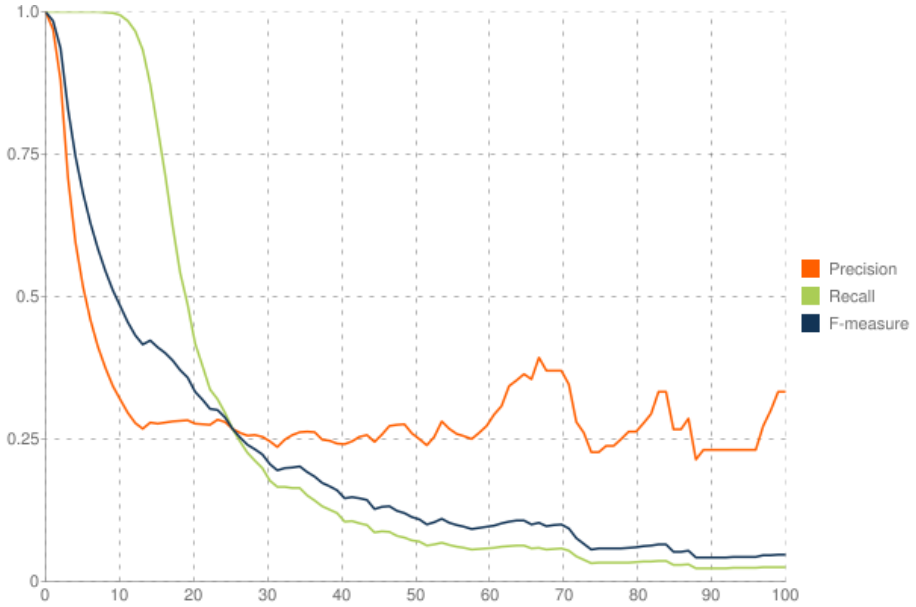


図 5.7: 実験 3：トレーニングデータの数を 1000 とした場合の精度

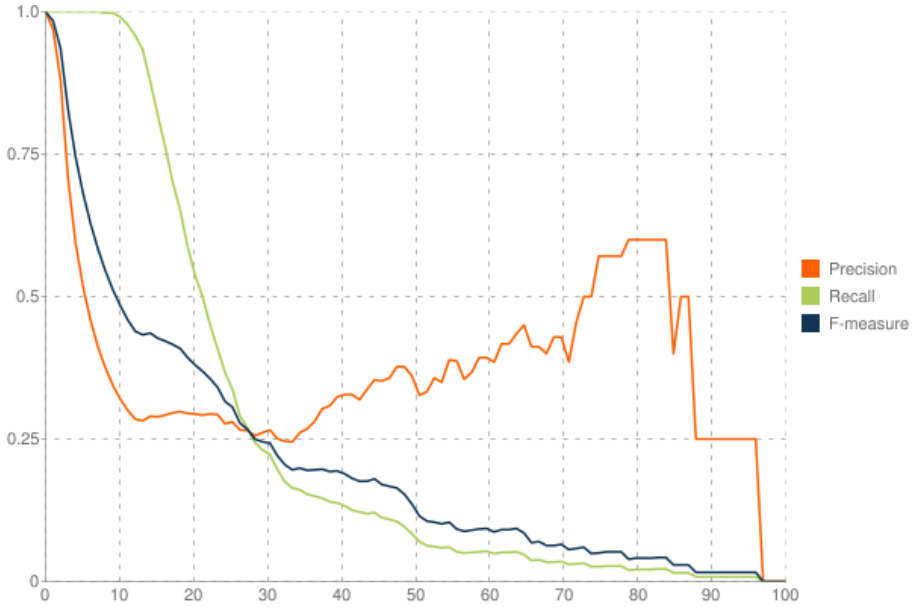


図 5.8: 実験 3：トレーニングデータの数を 2500 とした場合の精度

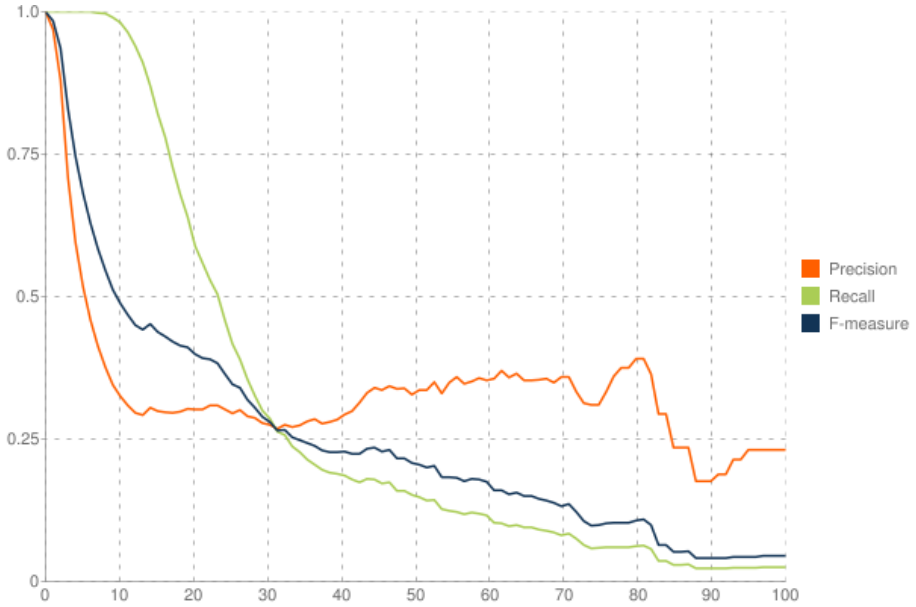


図 5.9: 実験 3：トレーニングデータの数を 5000 とした場合の精度

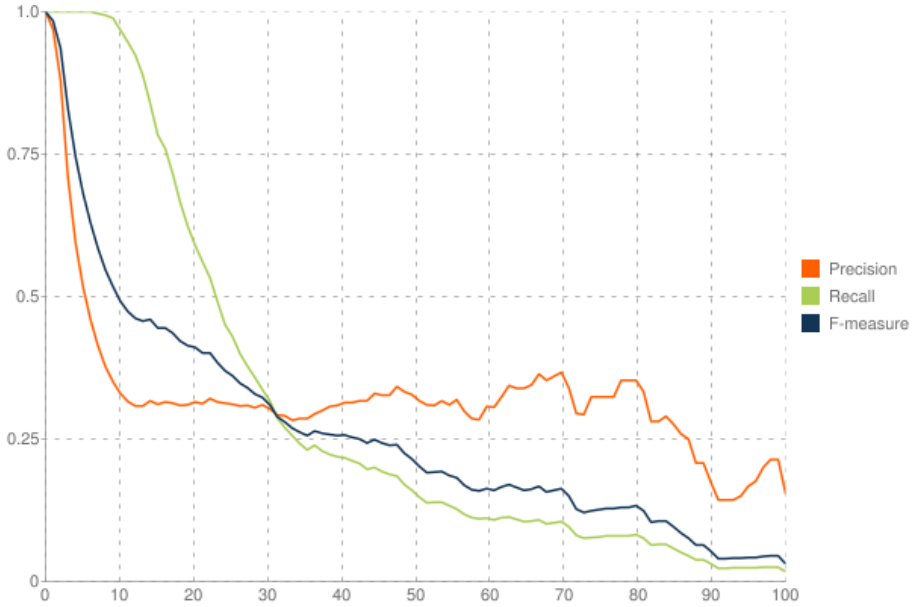


図 5.10: 実験 4：本文の単語出現の特徴量を抜いた場合の精度

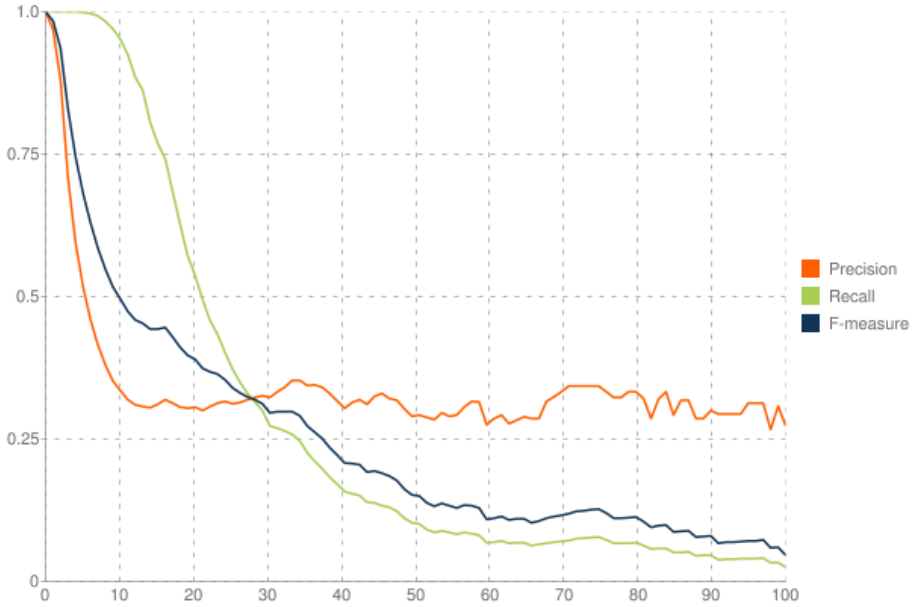


図 5.11: 実験 4：タイトルの単語出現の特徴量を抜いた場合の精度

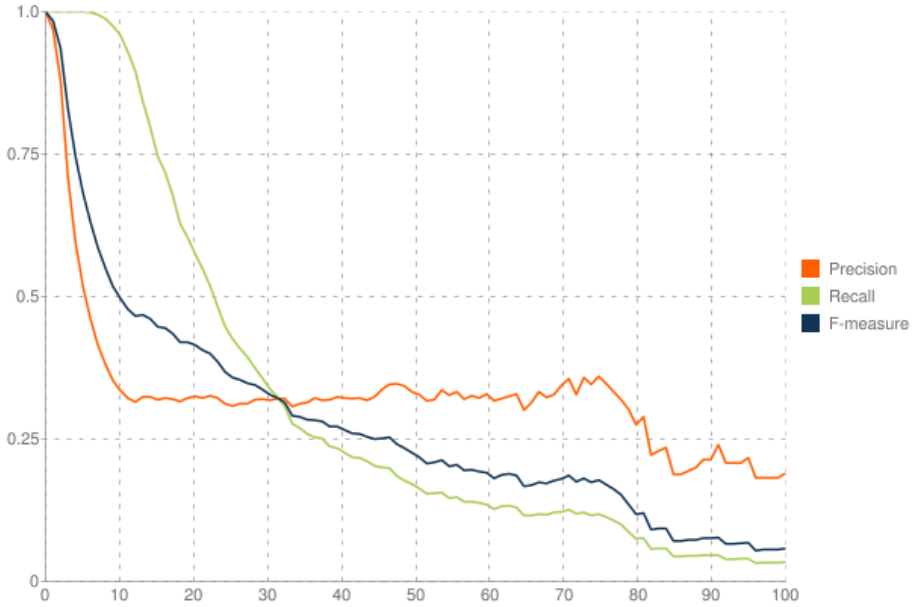


図 5.12: 実験 4：文字数の特徴量を抜いた場合の精度

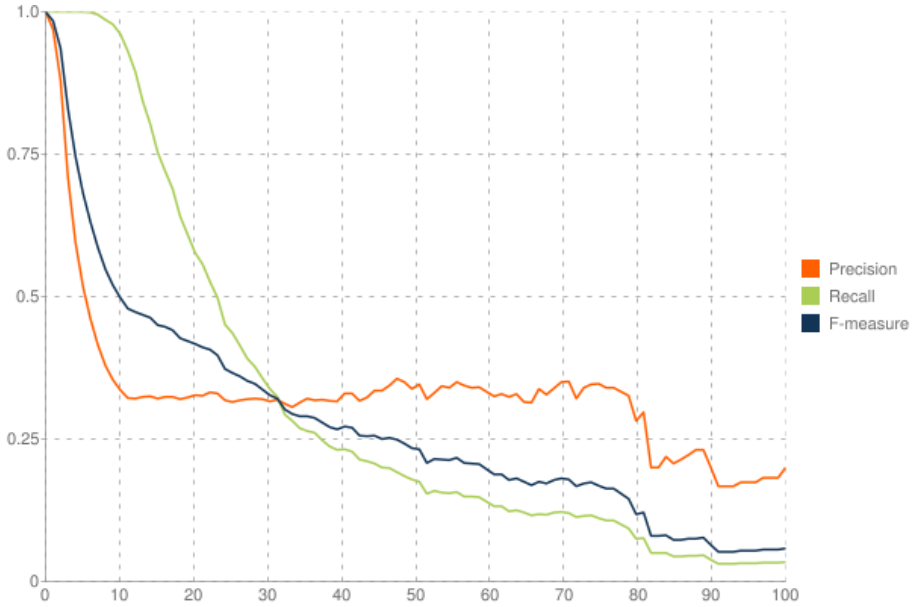


図 5.13: 実験 4：画像数の特徴量を抜いた場合の精度

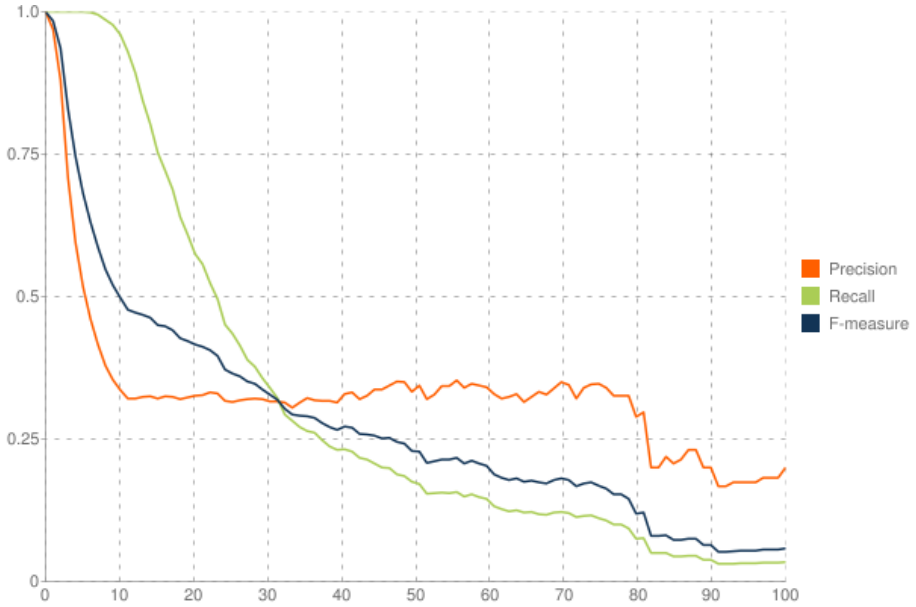


図 5.14: 実験 4：リンク数の特徴量を抜いた場合の精度

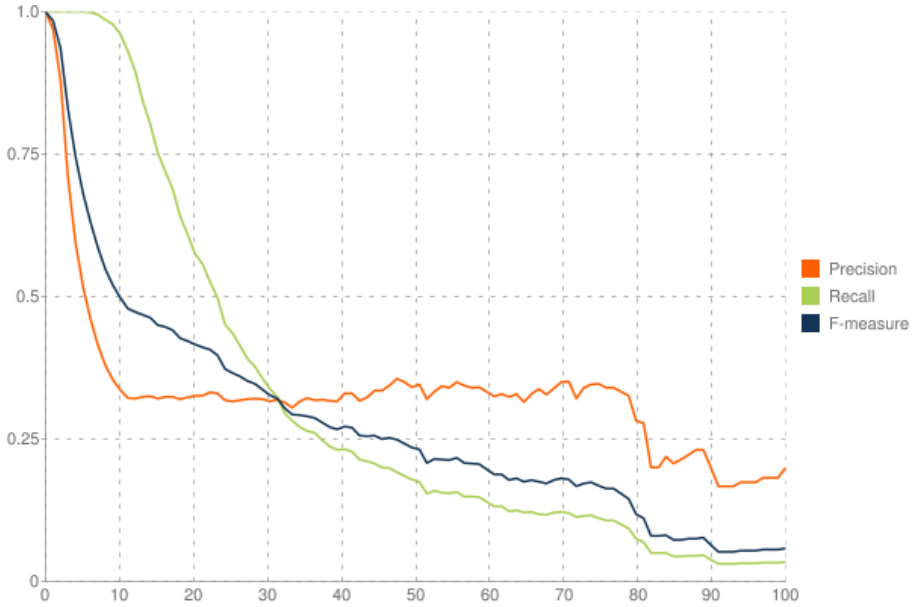


図 5.15: 実験 4：リンクの文字数の特徴量を抜いた場合の精度

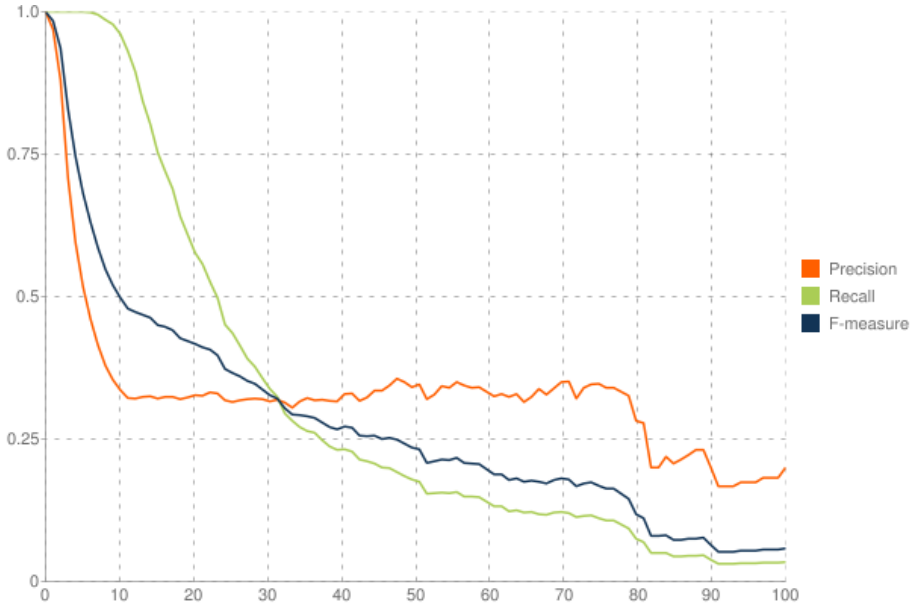


図 5.16: 実験 4：改行数の特徴量を抜いた場合の精度

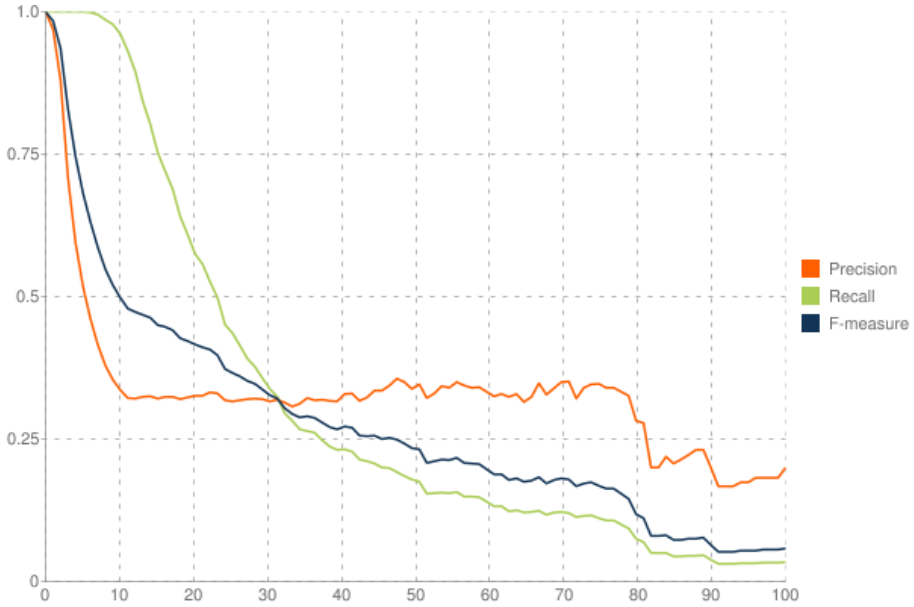


図 5.17: 実験 4：文字種類の特徴量を抜いた場合の精度

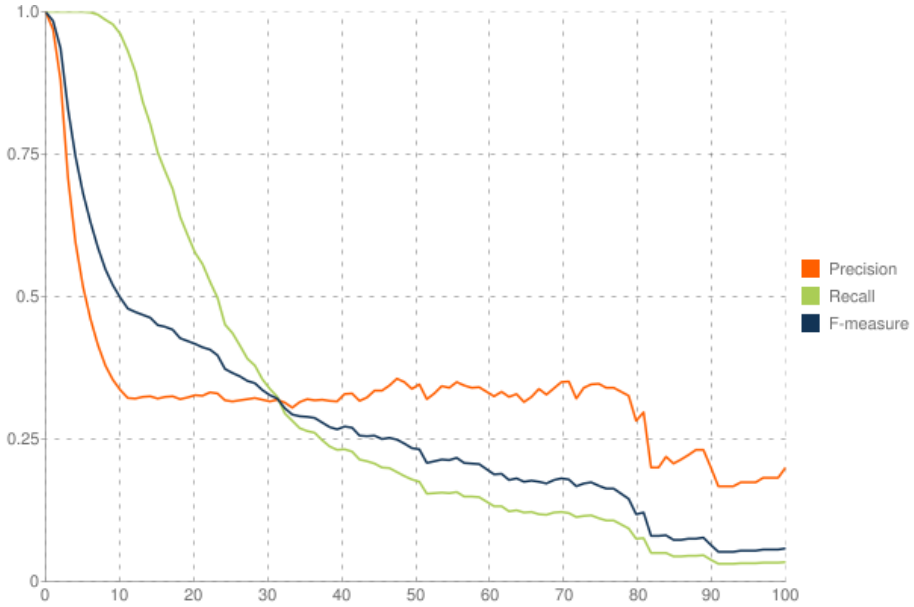


図 5.18: 実験 4：流行語の特徴量を抜いた場合の精度

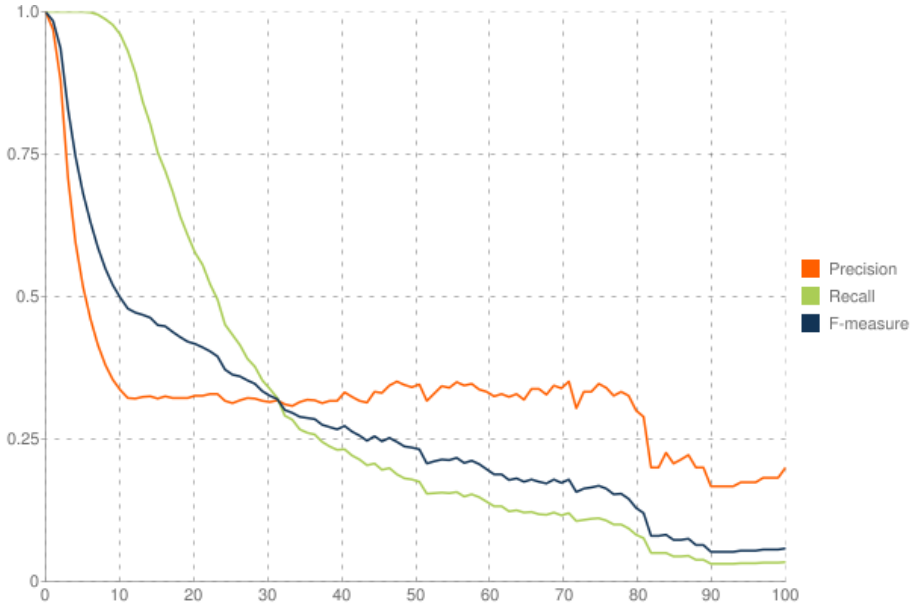


図 5.19: 実験 6：ページ全体のテキストを対象とした場合の精度

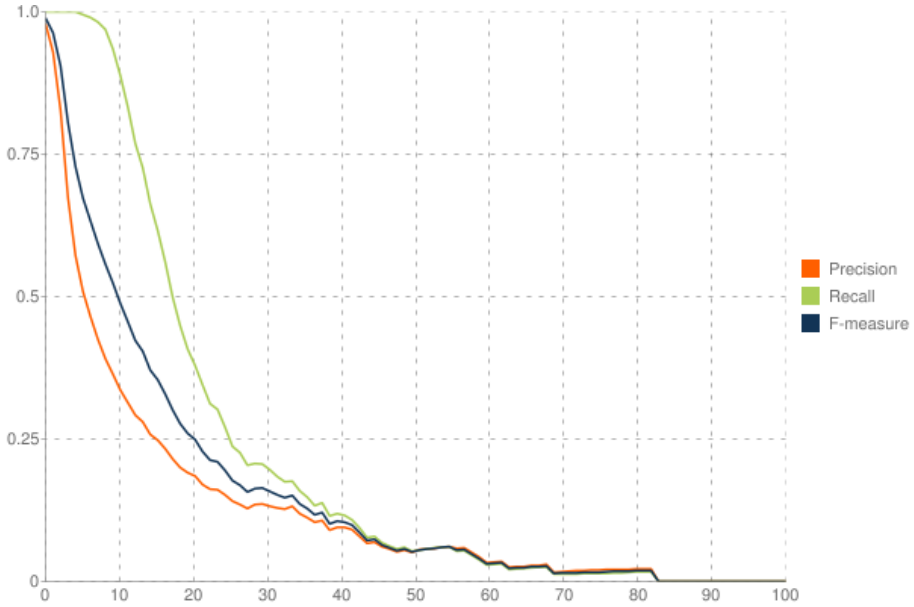


図 5.20: 実験 6：本文のみを対象とした場合の精度