G-13

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

Constructing regression models for predicting the numbers of social bookmarks

高松 征賢†　　鶴岡 慶雅†
Seiken Takamatsu　　Yoshimasa Tsuruoka

1．はじめに

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技術は 非常に重要なものとなっている。従来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(被リンク数の多い)Webページが重要なページとみなされる。そのため、Googleをはじめとした既存の検索エンジンでは、既に多くの人に知られているWebページが上位にランキングされる[1]。しかし、インターネット上には人に知られていなくても有用なWebページは大量にあり、それらを自動的に発見する技術は重要である。

本研究では、被リンク数などの外的要因を使わず、Webページの文章の構造などを特徴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。これによって、人に知られていないが有用であるWebページを自動的に発見することが可能になると期待される。

2．関連研究

ユーザがWebページにタグやコメントと共にブックマークを付けることができるソーシャルブックマークサービスが有用な情報源として注目されている。ソーシャルブックマークを使ったWebページ推薦に関する研究はこれまでにも行われてきたが[2][3]、本研究の目的は、ソーシャルブックマーク数の予測モデルを構築し、有用なWebページのランキングを作ることである。Golderら[4]は、ページが最初に投稿されてから一日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。毛受ら[5]は、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。Bleiら[6]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[7]のデータを使用して、Webページの人気を予測している。Hongら[8]は、twitter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。根本ら[9]は、非リンク数は使わずにソーシャルブックマークを付けたユーザ間の評価などを使用し、Webページの評価を行っている。高橋ら[10][11]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。

本研究では、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上記の研究と共通であるが、分類問題ではなくWebページの文章の構造などからブックマーク数の予測を行うという点で異なっている。

3．評価実験

本研究の提案手法の有効性を確認するために評価実験を行う。評価実験として、Support Vector Regression（SVR）で機械学習を行い、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う。SVRとは、学習モデルの一つであるSupport Vector Machineの回帰分析への拡張であり、カーネル関数を使うことで非線形回帰も行うことができる。

データとして、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[12]の総獲得ブックマーク数が上位1000位以内のblogから適当に20のblogを選び、それぞれ新着記事50件を抽出した合計1,000記事を使用した。ソーシャルブックマークのデータとして2011年7月29日に収集したはてなブックマークのデータを使用した（抽出した記事に付けられた合計ブックマーク数は71,772）。学習データとして、全体の5分の4に当たる800記事、テストデータとして残り5分の1に当たる200記事を使用した。特徴として、記事の総文字数、単語出現、画像の数、リンクの数、リンクの文字数、改行の数、文数、RSS購読者数、同じblogの新着記事50件に付けられたブックマーク数の合計を記事数50で割ったもの使用した。記事の総文字数は、短い文章よりも長い文章の方が有用であろうという考えから特徴として使用した。単語出現とは、記事の本文に形態素解析を行い名詞、形容詞、感動詞、動詞の単語を抽出し、各々の単語が本文内に出現した場合には0.1、しない場合には0の特徴量をあたえたものであり、ある特定の単語が出現した場合にブックマークが付けられるのではないだろうかという考えから特徴として使用した。リンクの数とリンクの文字数は、あまりにもリンクが多いとスパムである可能性が高いであろうという考えから特徴として使用した。最後の2つの特徴は本研究の目的である「人に知られていないが有用であるWebページを発見する」に反するが、2つの特徴と他の特徴とを比較するために使用した。

評価方法として、実際の値（ブックマーク数）と予測された値の差の絶対値の平均をとった。各特徴を除いた場合

† 北陸先端科学技術大学院大学 情報科学研究科, School of Information Science, Japan Advanced Institute of Science and Technology

の予測誤差を比較することにより、各特徴の有効性を調べる。

実験を行ったところ、すべての特徴を使用した場合、48.73となった。予測を行わない場合（予測結果をすべて0として評価した場合）は57.76となった。各特徴を除いた場合の予測誤差は表1のようになった。実験結果から、ブックマーク数の平均を除いた場合に予測誤差が一番広がっているので、ブックマーク数の平均が特徴として一番有効であるということがわかる。単語出現を除いた場合に予測誤差が縮まっているので、単語出現は特徴として有効ではないということがわかる。

表1 各特徴を除いた場合の予測誤差

特徴	予測誤差
記事の総文字数	48.70
単語出現	45.93
画像の数	48.72
リンクの数	48.72
リンクの文字数	48.72
改行の数	48.76
文数	48.67
RSS購読者数	49.00
ブックマーク数平均	52.38

4. おわりに

本論文では、Support Vector Regressionモデルでソーシャルブックマーク数を予測することにより有用であるWebページを発見する手法を提案した。また、各特徴を除いた場合の予測誤差を調べることで、各特徴の有効性を調べた。

今後の課題として、Webページの文章の構造だけではなく他の要素も考慮し、より精度の高い予測モデルの構築を試みる。

はてなブックマークのサイト内にあるランキングに乗るとソーシャルブックマーク数が増加する傾向があることから、そこに表示されたかどうかを考慮し予測モデルの構築を行う。

Webページにはそれぞれ特性がある。例えばプログラミング入門記事のようなレファレンスとしてあとで使うためにブックマークされたWebページや異論をコメントするためにブックマークされたWebページ、スパムのためにブックマークされたWebページなどといったようなものである。それらを分類し、それぞれを別々の問題として扱うことで予測モデルの精度向上を試みる。

今回実験で使用した特徴以外の特徴を探し有効性を調べることも重要な課題のひとつである。

さらに、本研究で提案した手法が実際に有用であるかどうかを判定するために、自動的に発見されたWebページ郡を第三者に見てもらい有用かどうかを主観的に判定する実験を行う予定である。

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] 佐々木, 宮田, 稲積, 小林, 酒井. Social Bookmarkにおけるコンテンツクラスタ間の類似度を用いたwebコンテンツ推薦システム. 情報処理学会論文誌, 47(20), pp. 14-27, 2007.

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

[5] 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.

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

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

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

[9] 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.

[10] 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.

[11] 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.

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