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

高松 征賢 (1010038)

北陸先端科学技術大学院大学 情報科学研究科

2012 年 2 月 6 日

キーワード: 機械学習、回帰分析、予測、ソーシャルブックマーク.

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。

本研究の提案手法の有効性を確認するため、実際のソーシャルブックマークのデータを用いて評価実験を行なった。回帰モデルとして、Support Vector Regression（SVR）を利用し、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なった。データとしては、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 [2] の新着エントリーに掲載された上位 10,000 件の Web ページを 2012 年 1 月 1 日から 2012 年 1 月 7 日の間に一日一回クローリングを行い取得した。また、ソーシャルブックマークのデータとしては、はてなブックマークの API を使用し、個々のページに付与されたブックマーク数を取得した。Web ページに、はじめてブックマークが付けられた時間から一週間後のブックマーク数を予測対象の値として使用した。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 件を使用した。

予測に用いる特徴量としては、特定の単語が Web ページ中、タイトル中に出現したかどうか、文字数、画像数、リンク数、リンクの文字数、改行数、文字種類、同一ドメインかどうか、タイトル中に流行語がいくつ含まれるかを使用した。

評価基準としては、予測誤差の平均および、ブックマーク数が閾値以上付与されるページを検出する精度（適合率、再現率、F 値）を用いた。実験を行なった結果、すべての特徴量を使用した場合には、対数領域での予測誤差が 0.9603 となった。ブックマーク数が 50 かどうかを予測した適合率は 34.1%、再現率は 17.9%、F 値は 23.5%となった。また、特徴量を種類ごとに削除して比較実験を行い、各特徴量の有効性を調べた所、ページ中の単語出現の特徴量が最も有効な特徴量であるという結果が得られた。以上の実験結果から、本研究で提案する手法が有効であることがわかった。

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、ユーザが読みやすいようなユーザインタフェースの設計をしているかどうかや、文章の簡潔さ、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の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。

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] はてなブックマーク. http://b.hatena.ne.jp/.