Constructing regression models for predicting the numbers of social bookmarks

Seiken Takamatsu (1010038)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 6, 2012

Keywords: Machine learning, Regression analysis, Prediction, Social bookmarks.

Today, the amount of information that exists on the Internet is huge. Information retrieval technology, which help people find desired information from the Internet, has become very important. Information retrieval systems such as Google usually consider Web pages that are known by many people, i.e., linked from many other pages, as important[1]. However, a problem with this kind of approach is that it is not useful in helping people to find useful Web pages that are not yet known to people, while there are many such unknown useful pages on the Internet.

Meanwhile, social bookmarking services are attracting attention as a means of managing and discovering useful Web pages as a community. Those services allow users to manage and share bookmarks with other unspecified users online. Users can also add additional information such as tags or comments to the bookmarks. Social bookmarks are a useful source of information because they are carefully selected by hand, and the number of bookmarks that each page has attracted can be regarded as an indicator of whether it is a useful Web page or not. However, social bookmarking has the same problem as in information retrieval—it is not useful in supporting users to discover useful Web pages that are not yet known by many people.

In this thesis, we propose a method for predicting the usefulness of a Web page solely from its content, i.e. without relying on external information such as the numbers of bookmarks or inbound links. More specifically, we use, as features for a regression model, words and characters in the text and the number of outgoing links. Our system is also expected to be useful in improving the ranking accuracy of search engines and in helping people to foresee the popularity of their newly created web pages.

To evaluate the effectiveness of our method, we conducted experiments using actual social bookmarking data. We used a Support Vector Regression (SVR) model as the

regression model to predict the number of bookmarks that each article has received during a week since its creation. To create the training and test data, we crawled the articles that appeared in the top 10,000 list of Hatena Bookmark[2] during the first week of January in 2012. The information on the numbers of social bookmarks was retrieved using the Hatana API. The numbers of social bookmarks that each articles has received during the week just after it received its first bookmark were used as the prediction target for our regression model. We sorted the articles according to the time when they received their first bookmark. The latest 5,000 articles were used as the test data, and the next 10,000 articles were used as the training data.

The features used in our SVR model are as follows: words in the title or the body of the article, number of characters, number of images, number of links, number of characters of the link, number of new line, character type, whether the same domain, number of buzzwords were used.

To evaluate our model, we measured its average prediction error and its accuracy of predicting whether each article will receive certain numbers of bookmarks in terms of precision, recall and f-measure. The experimental results show that the average prediction error was 0.9603 (in the logarithmic scale). Over 50 bookmarks's precision is 34.1%, and recall is 17.9%, F-measure is 23.5%. We also examined the effectiveness of each feature and found that words in the body of the article was the most effective. These results demonstrate the effectiveness of the method proposed in this thesis.

A possible topic of future work is to incorporate other kinds of features. For example, we could try user interface's design, brevity of the sentence, expertise of contents, consistency of content, whether it is there is no other content or not, etc.

References

[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] Hatena Bookmark. http://b.hatena.ne.jp/.