

{tag}

{/tag}

International Journal of Computer Applications

© 2012 by IJCA Journal

Volume 50 - Number 21

Year of Publication: 2012

Authors:

Jaber Karimpour

Ali A. Noroozi

Somayeh Alizadeh

10.5120/7924-0993

{bibtex}pxc3880993.bib{/bibtex}

Abstract

Web spamming tries to deceive search engines to rank some pages higher than they deserve. Many methods have been proposed to combat web spamming and to detect spam pages. One basic method is using classification, i. e. , learning a classification model from previously labeled training data and using this model for classifying web pages to spam or non-spam. A drawback of this method is that manually labeling a large number of web pages to generate the training data can be biased, non-accurate, labor intensive and time consuming. In this paper, we are going to propose a new method to resolve this drawback by using semi-supervised learning to automatically label the training data. To do this, we incorporate Expectation-Maximization algorithm that is an efficient and an important algorithm of semi-supervised learning. Experiments are carried out on the real web spam data, which show the new method, performs very well in practice.

ences

Refer

- Caverlee, J. , Webb, S. , Liu, L. , Rouse, WB. 2009. A Parameterized Approach to Spam-Resilient Link Analysis of the Web. *IEEE Transactions on Parallel and Distributed Systems*. 20: 1422-38.
- Caverlee, J. , Liu, L. 2007. Countering Web Spam with Credibility-Based Link Analysis. In *Proceedings of the twenty-sixth annual ACM symposium on Principles of distributed computing (PODC '07)*. 157-166.
- Caverlee, J. , Webb, S. , Liu, L. 2007. Spam-Resilient Web Rankings via Influence Throttling. *21st IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. 1-10
- Gyongyi, Z. , Garcia-Molina, H. 2005. Web Spam Taxonomy. *First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb'05)*.
- Ntoulas, A. , Najork, M. , Manasse, M. ,Fetterly, D. 2006. Detecting spam web pages through content analysis. In *Proceedings of the 15th international conference on World Wide Web*. 83-92.
- Castillo, C. , Donato, D. , Becchetti, L. , et al. 2006. A reference collection for web spam. *SIGIR Forum*. 11-24.
- Liú, B. 2011. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer.
- Wang, W. , Zeng, G. Tang, D. 2010. Using evidence based content trust model for spam detection. *Expert Systems with Applications*. 37: 5599-606.
- Gyongyi, Z. , Garcia-Molina, H. , Pedersen, J. 2004. Combating Web Spam with TrustRank. In *Proceedings of 30th Intl. Conf. on Very Large Data Bases (VLDB'04)*. 576-587.
- Becchetti, L. , Castillo, C. , Donato, D. , Leonardi, S. , Baeza-Yates, R. 2006. Link-based characterization and detection of Web Spam. *2nd Int Workshop on Adversarial Information Retrieval on the Web (AIRWeb'06)*. 1-8.
- Liu, Y. , Cen, R. , Zhang, M. , Ma, S. Ru, L. 2008. Identifying web spam with user behavior analysis. In *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*. 9-16.
- Erdelyi, M. , Garzo, A. ,Benczur, AA. 2011. Web spam classification: a few features worth more. In *Proceedings of the 2011 Joint WICOW/AIRWeb Workshop on Web Quality*. 27-34.
- Mitchell, T. 1997. *Machine Learning*. McGraw-Hill.
- Yahoo Research. 2007. *Web Spam Collections*, <http://barcelona.research.yahoo.net/webspam/datasets/>, accessed May 2011.
- Castillo, C. , Donato, D. , Gionis, A. , Murdock, V. , Silvestri, F. 2007. Know your neighbors: web spam detection using the web topology. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 423-30.
- Han, J. , Kamber, M. , Pei, J. 2011. *Data Mining: Concepts and Techniques*. Elsevier.

Index Terms

Computer Science

Information Sciences

Keywords

Adversarial Information Retrieval Web Search Web Spam Detection
Semi-supervised Learning

Expectation Maximization Algorithm