عنوان مقاله:

A Novel Set of Contextual Features for Web Spam Detection

محل انتشار:

مجله آنالیز غیر خطی و کاربردها, دوره 11, شماره 1 (سال: 1399)

تعداد صفحات اصل مقاله: 19

نویسندگان:

Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran - - -

Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran - - -

Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran - - -

خلاصه مقاله:

Web spam is one of the significant problems facing search engines. It wastes sources and time, decreases the quality of results and leads to user discontent. The two main approaches to the detection spam web pages are link and content-based analysis. In this study, we mainly focus on content-based analysis in both user-visible text and the source code of a web page to propose a set of features for web spam detection. we explore the relationship between types and frequency of HTML (HyperText Markup Language) tags used in a web page source code. We also examine the structure of the URL as the other source of information. Finally, the content of a web page visible to the user is considered semantically in order to identify relevance among the number of the existing topics in the text as well as the coherence of a text using Latent Dirichlet Allocation. Experimental results show that the proposed features increases the index of balanced accuracy from ۰.۳۳ to ۰.۶۹ and improves the web spam detection rate.

کلمات کلیدی:

web spam, content-based features, URL structure, HTML tags, topic modeling, Latent Dirichlet Allocation

لینک ثابت مقاله در پایگاه سیویلیکا: