ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 20 МАТЕМАТИКИ Выпуск 4
2013

А. А. Печников (Петрозаводск, ИПМИ КарНЦ РАН)**. О вебометрическом индикаторе «размер сайта».**

В проекте «Вебометрический рейтинг научных учреждений России» [1], как и практически во всех проектах, посвященных ранжированию веб-ресурсов, одним из индикаторов, характеризующих веб-сайт, является его размер [2, 3], для измерения которого, как правило, используют поисковые системы. В проекте [1] используется также программа BeeCrawler [4]. В процессе измерений вебометрических индикаторов иногда возникают ситуации, которые мы называем ошибкой поисковой системы. Пример ошибки Google: при измерениях количества страниц сайта Библиотеки по естественным наукам РАН на запрос в Google и Яндексе вида site:www.benran.ru, результаты SGoogle = 177000, SYandex = 6398, и почти тридцатикратная разница в результатах наводит на размышления об ошибке в Google.

При этом BeeCrawler в процессе сканирования создает для каждого сайта таблицу количества страниц на каждом уровне (до 7-го включительно). Ошибочные значения BeeCrawler практически в каждом случае могут быть объяснены наличием так называемых «паучьих ловушек», зацикливающих работу краулера. Ошибки Google и Яндекса объяснить невозможно в силу отсутствия необходимой информации о процессах сканирования в поисковых системах.

Для «сглаживания» ошибок Google и Яндекса предлагается подход, использующий значения BeeCrawler и гипотез относительно организации процессов сканирования в Google и Яндексе, для чего в каждом случае строится функция специального вида, позволяющая вычислить «правильные» значения индикаторов вместо ошибочных.

Проведены эксперименты с реальными данными, собранными для 400 сайтов РАН, что позволяет использовать точно процедуры исправления очевидных ошибок вместо слабо формализуемых мнений экспертов, в качестве которых пока выступают сами разработчики проекта.

Работа выполняется при поддержке Российского гуманитарного научного фонда (грант № 12-03-12001).

СПИСОК ЛИТЕРАТУРЫ

1. Вебометрический рейтинг научных учреждений России. URL: http://webometrics-net.ru.
2. Ranking Web of World universities. URL:http://www.webometrics.info.
3. *Антопольский А. Б.*, *Поляк Ю. Е.*, *Усанов В. Е.* О российском индексе веб-сайтов научно-образовательных учреждений. Информационные ресурсы России, 2012, № 4, с. 2–7.
4. *Печников А. А.*, *Чернобровкин Д. И.* Адаптивный краулер для поиска и сбора внешних гиперссылок. Управление большими системами, 2012, в. 36, с. 301–315.