ЗНАЧЕНИЕ ФАЙЛА ROBOT.TXT ДЛЯ ПОИСКОВЫХ СИСТЕМ И ПРАВИЛА ЕГО НАПИСАНИЯ

Дедушев О.В., Костин В.И.
Донецкий Национальный Технический Университет

1 Правила написания файла robot.txt

Одним из основных способов найти информацию в Internet являются поисковые машины. Поисковые машины, каждый день, исследуя “сеть” посещают веб-страницы и заносят информацию о них в базы данных. Это позволяет пользователю, набрав некоторые ключевые слова, увидеть, какие страницы удовлетворяют его запросу.

Понимание того, как работают поисковые машины просто необходимо вебмастерам. Для них жизненно важна правильная с точки зрения поисковых машин структура документов и всего сервера или сайта. Без этого документы будут недостаточно часто появляться в ответ на запросы пользователей к поисковой машине или даже вовсе могут быть не проиндексированы.

При индексации страницы роботом обращение происходит к файлу robot.txt, который должен присутствовать на каждом сервере. Этот файл описывает права доступа для поисковых роботов, причем существует возможность указать для различных роботов разные права. Для него существует стандарт под названием Standart for Robot Exclusion.

Файл /robots.txt предназначен для указания всем поисковым роботам (spiders) индексировать информацию сервера так, как определено в этом файле. Поисковые сервера всегда перед индексацией ресурса ищут в корневом каталоге домена файл с именем "robots.txt" (http://www.mydomain.com/robots.txt)[1].

Файл начинается со строки User-Agent, в которой описывается каким или какому поисковому роботу эта запись предназначается:

User-agent: googlebot

В данном случае обращение происходит к google.

В следующей строке: Disallow описываются не подлежащие индексации пути и файлы. Например следующая директива запрещает паукам индексировать файл email.htm:Disallow: email.htm

Директива может содержать и название каталога:

Disallow: /cgi-bin/

Эта директива запрещает индексировать каталог "cgi-bin".

В директивах Disallow[1] могут также использоваться и символы подстановки. Стандарт диктует, что директива /bob запретит паукам индексировать и /bob.html и /bob/index.html.

Если директива Disallow будет пустой, это значит, что робот может индексировать ВСЕ файлы. Как минимум одна директива Disallow должна присутствовать для каждого поля User-agent, чтобы robots.txt считался верным. Полностью пустой robots.txt означает то же самое, как если бы его не было вообще.

Любая строка в robots.txt, начинающаяся с #, считается комментарием. Стандарт разрешает использовать комментарии в конце строк с директивами, но это считается плохим стилем:

Disallow: bob #comment

Такая запись может привести к запрету на индексацию ресурсов bob#comment. Следовательно, комментарии должны быть на отдельной строке.

Пробел в начале строки разрешается, но не рекомендуется.

Disallow: bob #comment

Следующая директива разрешает всем роботам индексировать все ресурсы сайта, так как используется символ подстановки "*".

User-agent:*

Disallow:

Эта директива запрещает всем роботам это делать:

User-agent:*

Disallow: /

Данная директива запрещает всем роботам заходить в каталоги "cgi-bin" и "images":User-agent:*

Disallow:/cgi-bin/

Disallow: /images/

Данная директива запрещает роботу googlebot индексировать файл cheese.htm:

User-agent:googlebot

Disallow: cheese.htm

2.Почему необходимо писать robot.txt

Поисковые машины индексируют все без исключения файлы, которые лежат в незащищенном доступе также. Винить в этом необходимо в первую очередь разработчиков, которые размещают информацию совершенно не задумываясь о последствиях. Не пишут файлы robot.txt или пишут их неправильно, причем поисковик не специально показывает такую информацию

3. Способы повышения эффективности поиска

Если правильно сформулировать запрос поиска, в считанные секунды можно получить интересную информацию. Примером может послужить следующее: многие компании в последнее время преподносят презентации, созданные в программе PowerPoint, всем известно что такие файлы имею расширение ppt. Для того, чтобы было понятно что информация предназначена для узкого круга адресатов, каждый слайд снабжается пометкой «Конфеденциально» или «Для служебного использования». Для нахождению такой информации вводится такой запрос[2]:

Ext:ppt confidential «for internal use only»

Меньше одной секунды и найдены документы принадлежащие компании siemens, cisco, sun. С помощью оператора ext: можно целенаправленно искать ссылки на какой либо файловый формат. В качестве альтернативы можно попробовать и “filetype”.

С помощью безобидного кода можно найти ссылки которые ведут к информации, необходимой для входа в профиль : Inurl: «login.asp».

Allintitle:[2] ограничивает результаты поиска только теми страницами, в строке заголовка которых встречаются все параметры поискового запроса.

Link:[2] - набрав перед запросом это оператор , вы найдете все сайты встречающиеся на указанную страницу.

При правильном построение запроса имеется возможность найти много полезной информации, которую нельзя найти, если задавать слова при поиске.

Литература

[1] J Long. Google Hacking.Syngress Publishing.Inc,2005.

[2] Интернет ресурс http://www.nnm.ru.