Ранжирование источников в корпоративной системе интеграции и мониторинга новостей

Ландэ Д.В., Григорьев А.Н., Брайчевский С.М., Дармохвал А.Т.
Информационный центр «ЭЛВИСТИ», Киев, Украина

Мощные информационные возможности Интернет порождают проблему оптимизации состава и количества источников, используемых корпоративной системой интеграции и мониторинга новостей для обеспечения приемлемого качества ее функционирования. В этой связи, актуальными оказываются вопросы ранжирования источников новостной информации. Принципам ранжирования отдельных веб-документов посвящено большое количество работ, вместе с тем, вопросам ранжирования и отбора веб-сайтов с учетом их контента, объемов и стабильности тематики публикаций существенного внимания до сих пор не уделялось.

Известно, что распределение источников по контенту, соответствующему тематическим потребностям корпоративного пользователя, удовлетворяет закону Бредфорда, соответственно, при отборе источников обязательно должно учитываться их ранжирование по степени соответствия тематике. Кроме того, представляется перспективным дополнить традиционный подход более объективными и строгими методами, позволяющими оптимизировать процесс формирования информационной базы корпоративной системы.

В качестве экспериментального корпуса в исследованиях авторов использовался информационный массив, охватываемый системой интеграции и мониторинга новостей InfoStream, в частности, информационный массив за март 2008 года объемом свыше 1.2 млн. документов из более чем 2500 веб-сайтов. Исследовалась зависимость количества документов, опубликованных источниками, ранжированными по «производительности». Были получены результаты, свидетельствующие о близости рассматриваемой зависимости к гиперболической (т.е. о действии обобщенного закона Ципфа). Эти результаты позволили построить критерий отбора необходимой для корпоративных применений части источников из общего объема источников, охватываемых системой InfoStream. Если предположить, что все источники давали бы одинаковый вклад по количеству опубликованных документов, то зависимость количества документов в системе от количества источников была бы линейной (f_{lin}) и выражалась бы формулой:

$$f_{lin}(n) = n\frac{f_{\max}}{N},$$

где n - номер источника в ранжированном списке, $f_{\max}$ - максимальный объем охватываемых документов, N - количество источников.

Очевидно, что отклонение реальной зависимости от линейной сначала возрастает, а затем уменьшается до нуля. Как критерий оптимального выбора рассматривается пороговое количество источников n_p, когда значение реальной зависимости $f(n)$ максимально отклоняется от приведенной линейной:

$$n_p = \arg\max \ \{f(n) - f_{lin}(n)\}.$$

При этом количество охватываемых документов, соответствующих n_p при максимальном количестве источников (2500) достигает 80% от $f_{\max}$. Таким образом, построенная модель согласуется с принципом Парето: приблизительно 20% наиболее продуктивных источников публикуют 80% документов.

Предложенный количественный критерий отбора источников информации позволил расширить спектр применяемых ранее качественных методик отбора информационных источников для корпоративных применений системы интеграции и мониторинга новостей InfoStream.