Рейтинг онлайн-СМИ на основе дублирования новостей

Тезисы доклада на семинаре «Поисковые технологии 2010»

Антонов Александр, Баглей Станислав, Ландэ Дмитрий

alexa@galaktika.ru, baglei@galaktika.ru, dwl@visti.net

1. Введение

Появление в Вебе достаточно большого выбора среди онлайн-СМИ вызвало необходимость оценки их достоверности. Сравнительно небольшое количество наиболее крупных источников информации находятся "на слуху" у пользователя, о качестве большинства прочих на текущий момент ему остается только догадываться. Рейтинг онлайн-СМИ в этой ситуации может быть средством как явного вспомогательного инструмента – информации, предоставляемой для просмотра, так и неявного (применения полученных данных для ранжирования новостных сообщений при показе новостей). В нашем случае целью было создание рейтинга онлайн-СМИ для агрегирующего новостного ресурса www.webground.su.

2. Обзор существующих подходов к составлению публикуемых рейтингов новостных источников

- Newsknife.com: по частотности появления ссылок на источник в сервисе Google News. Анализируется главная страница сервиса: рассчитывается статистика попаданий в две лучшие новости сюжета, в лучшую новость сюжета на странице всех новостей сюжета, на основе лучшего распределения в топ-7 позициях. Отдельно анализируются лучшие новички среди новостных ресурсов; в 2009 году с большим отрывом таким ресурсом стала Википедия;
- Webscan: анализ текстового контента Рунета. В рейтинг попадают наиболее часто упоминаемые бренды. Наивысшие показатели – у брендов новостных ресурсов;
- Медиалогия: анализ текстового контента новостных веб-ресурсов. При составлении рейтинга учитывается количество ссылок на сообщение и заметность цитирующих сообщений. Заметность сообщения зависит от авторитетности СМИ, месте и объема размещения публикации, наличия и размера сопроводительных иллюстраций, эмоционального окраса заголовка и прочих параметров;
- Рейтинг СМИ в Яндекс-блогах на основе дневников пользователей Сети;
- Hitwise.com, Comscore.com, Nielsen.com: по аудитории СМИ, рассчитываемой на основе анализа трафика посетителей. Оценивается доля рынка, количество посещений, количество уникальных посетителей;
- Reddit.com, Digg.com: интерактивный режим формирования рейтинга новостей пользователями. Посетители, поодиночке и в сообществах, голосованием повышают интересные новости и понижают неинтересные;
- один из первых проектов рейтингов СМИ в Рунете - совместный проект СМИ.ru и Фонда эффективной политики Глеба Павловского. Уже несколько лет существование сервиса не поддерживается.

Использовать для нашей задачи какой-либо из рассмотренных рейтингов было невозможно либо из-за недостаточной полноты публикуемых данных, либо из-за закрытости

алгоритмов, на базе которых строятся рейтинги, вследствие чего были неясны четкие критерии их потенциальной применимости. Кроме того, в нашем распоряжении были уникальные данные мониторинга новостных сайтов, которые позволили нам разработать собственный подход к составлению рейтинга.

3. Алгоритм формирования рейтинга новостных источников для ресурса webground.su

Ранжирование источников основано на информации о группах найденных новостей-дубликатов и признаком времени публикации, приписанном новостям.
На первом этапе алгоритм поиска дублирующихся сообщений разбивает множество новостных сообщений на непересекающиеся подмножества. После чего в каждом подмножестве сообщения ранжируются по времени публикации в убывающем порядке. Каждое из выделенных подмножеств представляется в виде направленного графа, вершинами которого являются сообщения, а ребрами – отношения в упорядочении внутри подмножества.

С целью сокращения вычислительной сложности алгоритма принято ограничение, при котором ребра могут соединять только соседние элементы упорядочения. Каждое из ребер направлено от более раннего к более позднему сообщению.
К построенным графам применен алгоритм PageRank, с помощью которого каждой из вершин-сообщений на графе присваивается соответствующий вес. Использована версия алгоритма PageRank с ненормированными весами ребер.

Для составления итогового рейтинга источника учитывается накопленная месячная статистика веса PageRank сообщений источника и среднее время запаздывания при публикации новостей.

4. Результаты работы

Анализ результатов формирования рейтинга (пример: http://webground.su/sources.php?param=SourceList&Sort=2&Filter=0).

Проблемы формирования, потенциал применения, пути развития.