

УДК 681.3; 004:519.816; 519.688

ПРИНЯТИЕ РЕШЕНИЯ ПО ОПТИМИЗАЦИИ ПРОЦЕССА ДОСТАВКИ ИНФОРМАЦИИ В ИНТЕРНЕТ

А.С. Панчук

МНУЦИТС ЮНЕСКО/МПИ, Киев, Украина

e-mail: oleksander.panchuk@cbn-cis.org

Количество информации размещенной в сети Интернет увеличивается экспоненциально, – на 150%-200% ежегодно [1]. В связи с этим, эффективность доставки информации конечному пользователю стремительно падает. Даже важная информация, из-за возрастающей «зашумленности» не находит «своего» пользователя, а пользователь не находит искомую информацию.

Технологический процесс доставки информации в Интернет включает четыре последовательные стадии:

- 1) Инкапсуляция информации в носитель (создание Web страниц).
- 2) Реклама информационного материала (информирование потенциальных пользователей о ее содержании, с указанием мест размещения).
- 3) Передачи информации заинтересованным пользователям (публикация Web страницы).
- 4) Получение отклика на информацию от пользователя.

Три первые стадии отвечают за размещение информации в Интернет, что включает публикацию ее на Web страницах сайта с последующим размещением в базах данных поисковых серверов (ПС). Вторая стадия является ключевой в процессе доставки информации, и с каждым годом все более актуальной, трудно реализуемой и затратной.

Основную роль в поиске информации выполняет ПС, который на запрос пользователя обычно выдает более сотни тысяч страниц содержащих слова из запроса. Подавляющее большинство этой информации не релевантно целям поиска пользователя [2]. Соответственно собственник Web ресурса оказывается перед проблемой: каким же образом информировать пользователей из «целевой группы», target group, от которых автор этого материала ожидает – реакции. ПС, обычно, выдает информацию пакетами по десять ссылок на страницу. Распространено мнение, что массовый пользователь, в основном, при просмотре результатов выдачи ПС не идет далее третьей страницы, что соответствует первым 30 ссылкам [3]. Таким образом, чтобы пользователь смог найти информацию, ее владелец, находясь в условиях жесткой конкуренции и ограниченности средств, должен продвинуть Web ресурс в «тридцатку» первых в выдаче ПС.

ПС размещает в своей базе данных Web страницу индексируя ее. Исходя из общих представлений о работе ПС, можно утверждать, что для каждого информационного ресурса создается уникальный ранжированный список ключевых слов, keywords (КС). Будем подразумевать под КС также и их словосочетания. Каждое из этих КС имеет свой вес или ценность. Причем функция по которой вычисляется вес ключевого слова неизвестна конечным пользователям и отличается у разных ПС. Как правило она нелинейна и зависит от следующих параметров: относительная плотность КС на странице, количество внешних ссылок на Web страницу и их рейтинг, плотность КС в текстах размещенных на других страницах и ссылающихся на данную Web страницу, времени жизни данной страницы и пр. [2, 3, 5]. Пользователь ПС, как и автор ресурса не может получить доступ к значению веса этих КС. И фактически, когда в строку поиска вводится запрос, результатом его обработки является уже ранжированный по базе данных ПС список, состоящий из ссылок на все релевантные Web страницы. Позиция конкретной ссылки определяется ее весом по запрашиваемому КС, относительно других проиндексированных страниц.

Владелец ресурса может повлиять лишь на некоторые факторы от которых зависит вес страницы по конкретному КС. Часть факторов является внешними, как например, временные или ссылочные. Поэтому, в условиях жесткой конкуренции и ограниченности средств, продвинуть ресурс в первую тридчатку выдачи результата запроса по всем КС невозможно. И это одна из причин, почему ПС рекомендуют ограничиться списком из 10-15 слов или фраз, наиболее точно характеризующих Web-ресурс (например, рекомендации на рекламу [4]). То есть, по такому количеству КС можно достичь «реальной» его видимости и успеха в продвижении.

Этот уникальный список КС, называют – «семантическое ядро» (СЯ) сайта [3]. Чтобы попасть в первую тридчатку выдачи ПС по запросам большинства КС заявленных в СЯ, используется хорошо оснащенная и контролируемая процедура [5, 6, 7], носящая итерационный характер:

1. Экспертом составляется СЯ Web-ресурса.
2. С помощью программ для продвижения [5, 6], сайт анализируется на соответствие ресурса требованиям ПС и составляется список рекомендаций по улучшению Web-сайта. Соответствующим образом модифицируют структуру страниц, их наполнение, покупаются ссылки с тематически близких Web серверов (например, как это описано в [5]) и тому подобное.
3. Сайт подается на индексацию ПС.
4. С помощью программ для сбора статистики и анализа посещаемости [6, 7] определяется эффективность (себестоимость) в достижении поставленных целей, по каждому КС (т. е. их отдачу в плане привлечения пользователей из целевой группы). Модифицируется СЯ.
5. Процесс повторяют начиная с этапа 2, до получения желаемого результата.

Практически, эта процедура очень затратна и занимает не менее шести месяцев. Количество ее итераций и их длительность напрямую зависит от качества подбора исходного СЯ. Традиционно, подбор КС для СЯ осуществляют ранжированием по популярности списка альтернатив (для их выбора используют средства автоматического реферирования или экспертные методы), – слов и фраз, релевантных исходной информации и (или) желаниям владельца Web-ресурса [5, 6]. Из него оставляют первые 10-15 или немного более КС.

Наиболее популярные (запрашиваемые) слова, являются и более общими словами – релевантными ресурсам из большого класса различных тематических областей. Соответственно, больше конкуренция, больше случайных пользователей, а для целевой группы – поиск в среде с большим шумом. Продвижение по таким КС тяжелее и затратней с каждым годом. И как далее будет показано, в ряде случаев, их выбор относительно менее популярных КС, но более частных, существенно не прибавит пользователей из целевой группы. Если же представить рекламирование информации в виде процесса фильтрации, где фильтр «реклама» ограничен 10-15 КС, то традиционный подход демонстрирует низкую селективность и непрогнозируемость результата, – пропускается значительное число случайных пользователей, без возможности какой либо оценки прихода пользователей из целевой группы. Определяя эффективность фильтрации, как величину прямо пропорциональную количеству прошедших пользователей из целевой аудитории и обратно пропорциональную затracенным ресурсам, можно утверждать – традиционный подход ведет к построению не оптимальных фильтров и удорожанию всей процедуры продвижения.

Используются улучшения традиционного подхода. Сайт анализируют на конкурентно-способность с первой тридчаткой информационных ресурсов (не обязательно из родственной предметной области) выдаваемых ПС [3] и исключаются затраты на продвижение ресурса по КС недостижимым относительно средств и времени

выделенных на рекламу. Это уменьшает непроизводительные расходы, но не увеличивает долю пользователей из целевой группы.

Как итог сказанного сформулируем задачу: необходима методика прогнозирования и алгоритм выбора такого фильтра, изменения которого на протяжении всего итеративного цикла продвижения, будут минимальны. Такой фильтр можно понимать как оптимальное СЯ.

Определим целевую аудиторию (ЦА) – как тех пользователей, цель поиска которых совпадает с информацией представленной на Web-ресурсе и может быть выражена ограниченным набором КС состоящих из фраз конечной длины. Условимся называть корневыми ключевыми словами (ККС) те из них, которые однозначно и точно описывают размещенную информацию и включают другие КС рассматриваемого набора, но сами целиком, не являются частью какого либо из этих КС. Построить прогноз для выбора КС для СЯ возможно, рассматривая следующую модель поведения пользователя из ЦА, впервые ищущего информацию: поиск осуществляется посредством ввода все более усложненного запроса (КС), до тех пор, пока в результате выдачи ПС не встретится релевантный ресурс. Другими словами: процесс поиска информации пользователем из ЦА – это итерационный процесс пошагово сужающий область поиска посредством усложнения поискового запроса. Построим набор из траекторий поиска пользователей ЦА и назовем его набором деревьев целей (ДЦ). Допустим, что нет ни одного релевантного целям поиска пользователей ресурса, и они в качестве начального КС вводят фразу состоящую из одного слова. Затем, пользователи добавляют по одному уточняющему слову и в итоге приходят к необходимости вводить в строку поиска одно из ККС. Другими словами, каждый уровень отдельно рассматриваемого ДЦ, будет состоять из набора КС, включающих КС из верхнего уровня и размерностью на одно слово больше. В корне дерева будет уникальный ККС, максимальной размерности – КС предельно точно и однозначно описывающее искомый информационный ресурс и выражающее сформулированную цель поиска предполагаемого пользователя из ЦА. В данном контексте мы подразумеваем КС реально используемые пользователями при поиске информации (например, такие данные может получить эксперт, как дополнительную услугу на ПС: Google Suggestion, Yandex-Direct или wordtracker).

Назовем КС, на которых пользователь завершает поиск первый раз встречая интересующую его информацию, предельным ключевым словом (ПКС). Тогда, набор состоящий из всех КС включающих ПКС и ниже по ДЦ, с точки зрения владельца Web ресурса, является предметом конкуренции с тематически-родственными ресурсами. Назовем это множество эффективным набором КС (ЭНКС). Понятно, что КС, расположенные выше ПКС по дереву целей, являются более общими КС, родственными много более широкому классу ресурсов, и полному набору тематически-родственных рассматриваемому ресурсам. Эти КС будут более популярными, но гораздо более затратными для продвижения. Рассматривая ДЦ, можно показать, что количество пользователей из ЦА, вводящих это КС будет такого же порядка или меньше, чем для нижележащего, но более специализированного ПКС. Соответственно, область выше ПКС есть более затратная (так как здесь более широкая конкуренция), с минимальной отдачей и не продуктивная для приложения усилий по рекламе (пользователь ЦА, не найдя в первой тридцатке результатов выдачи ПС даже конкурентов, в любом случае, вынужден будет усложнять запрос до ввода ПКС).

Таким образом, мы можем сформулировать основной критерий отбора КС для построения СЯ Web-ресурса: с целью значительной экономии средств, и реального достижения пользователей из целевой аудитории необходимо использовать КС являющиеся предметом конкуренции. То есть основой СЯ будут все КС из ЭНКС.

Идея алгоритма выбора оптимального СЯ, основывается на первоначальном построении высокоселективного, низкозатратного фильтра «реклама» в качестве затравочного семантического ядра, состоящего исключительно из ККС. Статистика утверждает, что в среднем, пользователи используют КС состоящие из двух-трех слов [3]. Рациональным решением есть - использование в качестве ККС, фраз состоящих из четырех ключевых слов. Далее, итеративно, формируются более общие КС (строится ДЦ), которые проверяются на наличие конкуренции с ресурсами из тематическо-родственной области. Конечно, их включение уменьшит показатель селективности и увеличит показатель затратности, но позволит ресурсу не быть выброшенным из реальной конкурентной борьбы.

Запишем алгоритм более подробно:

1. Строим альтернативы и их синонимы как ключевые фразы состоящие из 4 слов. Для этого используем систему автоматического реферирования (для поиска прямых КС) и перевода (для поиска синонимически эквивалентных КС).
2. Оставляем из них те, которые отвечают требованиям для ККС. Это первый уровень набора ДЦ (создаем первый хэш-массив).
3. Строим следующий уровень ДЦ (как новый хэш-массив). Генерируем набор более общих КС, по порядку исключая одно слово, из каждого КС предыдущего уровня. Проверяем, каждое новое КС на наличие конкуренции из тематической области Web-ресурса. Если есть прямые конкуренты, запоминаем КС в хэш-массиве, иначе генерируем следующий элемент текущего уровня ДЦ, пока не переберем все КС из предыдущего уровня.
4. Повторяем шаг 3, до тех пор пока в генерируемых КС, количество слов не достигнет 1, включительно, или хэш-массив по завершению шага 3 не будет содержать ноль элементов.
5. Просматриваем все хэш-массивы на наличие КС не употребляемых пользователями для поиска и удаляем их.
6. Если общее количество КС больше 15, ранжируем на каждом уровне (начиная с корневого) КС по частоте включения слов из КС верхнего уровня и удаляем наиболее (полно) цитируемые, оставляя КС содержащие уникальные слова.

Описанный метод получения эффективного семантического ядра, гарантирует при его продвижении на ПС максимальную долю целевой аудитории в потоке посетителей при ограниченных затратах на рекламу.

Литература

1. How much Information? Peter Lyman, Hal R. Varian, 2000 Regents of the University of California, <http://www2.sims.berkeley.edu/research/projects/how-much-info/how-much-info.pdf>
2. Каким станет интернет-поиск в будущем? Наталья Евдокимова, Москва, 2007 http://internet.cnews.ru/reviews/index.shtml?2007/04/20/246848_1
3. Система анализа конкурентной среды в Интернете, Сергей Титков, Альманах "Лаборатория рекламы, маркетинга и public relations", Москва, № 3` 2005 г. http://lab.advertology.ru/frem_anons.htm
4. Adwords, help system, 2007, <https://adwords.google.com/>
5. Successful Search Engine Marketing, Axandra, 2007, <http://www.axandra.com/>.
6. WebTrends Analytics, WebPosition, <http://www.webtrends.com/>.
7. NetTracker, <http://www.sane.com/>.