АЛГОРИТМ ФОРМИРОВАНИЯ АССОЦИАТИВНЫХ СВЯЗЕЙ И ЕГО ПРИМЕНЕНИЕ В ПОИСКОВЫХ СИСТЕМАХ

В. И. Шабанов
Рамблер Интернет Холдинг
vs@rambler-co.ru
А. Е. Власова
Московский государственный лингвистический университет
a_vl@rambler.ru

В докладе представлен алгоритм построения ассоциативных связей на массиве запросов к поисковым системам в Интернете. Отличительной особенностью алгоритма является то, что связи строятся на основе анализа временных интервалов между поисковыми запросами. Алгоритм формирует списки запросов, поданных каждым из пользователей, выделяет в каждом из списков группы запросов, поданных в течение небольшого интервала времени, а затем производит объединение, ранжирование и фильтрацию полученных групп. Для ранжирования применяется косинусообразный коэффициент корреляции между списками ассоциативных связей, вычисляемый для каждой пары запросов.

Приведены примеры ассоциативных связей на массиве запросов к поисковой системе "Рамблер", продемонстрированы функциональные возможности использования ассоциативных списков при автоматической обработке запроса.

В качестве основных направлений использования "ассоциативного модуля" в докладе рассматриваются: уточнение / расширение пользовательского запроса; получение новых знаний на основе списка ассоциаций; планирование развития web-серверов; планирование рекламных компаний в Интернете, таргетинг и формирование целевой аудитории.

1. Легко ли найти информацию в Интернете?

Интернет постепенно стал частью нашей повседневной жизни. Количество сайтов постоянно растет, и все меньше становится людей, которые при слове "сеть" недоумевающе пожимают плечами. Интернет – это современное, удобное средство коммуникации и обширный информационный источник. В этой статье мы не будем останавливаться на вопросе, насколько "много" информации можно найти в сети, и какого качества эта информация. Попробуем оценить доступность информации для конечного пользователя.

Самый простой вариант информационного поиска в www-пространстве - когда пользователю известен уникальный сетевой адрес информационного источника (URL). В этом случае доступ к информации может быть ограничен только техническими параметрами, например, скоростью и надежностью связи.

Но в большинстве случаев пользователю приходится использовать специальные поисковые инструменты. Чаще всего в роли таких инструментов выступают поисковые системы и каталоги. Анализ массива запросов к поисковым системам (в частности, запросов от одного пользователя за время отдельной сессии работы с поисковой системой) показывает, что пользователю приходится неоднократно переформулировать запрос, прежде чем он получит список найденных документов, в котором содержится название и адрес искомого информационного источник.

Например, в поисковой системе "Рамблер" поиск российского сайта, посвященного известному иллюзионисту Дэвиду Копперфильду, может начинаться с запроса ***копперфильд***. После того, как требуемый сайт не был получен

в списке найденных документов (по крайней мере, на первой странице выдачи), запрос расширяется и переформулируется: ***дэвид копперфильд***, ***секреты копперфильда*** и т.п. И так до тех пор, пока в списке найденного не появится нужный сайт, или пока пользователю не надоедят бесплодные попытки.

Каков семантический механизм "расширения" и "дополнения" запроса? Многие пользователи используют различные парадигматические связи между словами для уточнения запросов, в частности, синонимы и варианты (например, замена запроса ***СВЧ-печь*** на ***микроволновка***). Часто к запросу добавляется слово (или несколько слов), являющееся обозначением более общего, родового понятия. Например, такая ситуация возникает при расширении запроса ***аквариум*** (может относится к предметам - аквариум для рыбок, - и к названию музыкального коллектива "Аквариум") с использованием более общего понятия "музыка" - ***аквариум музыка***.

К сожалению, не каждый пользователь способен самостоятельно переформулировать запрос таким образом, чтобы это привело к повышению релевантности найденных документов. Ниже рассматривается алгоритм, позволяющий пользователю с минимальными временными потерями уточнить запрос и получить нужную информацию.

2. Что такое ассоциативные запросы?

В общем виде под "ассоциациями" или "ассоциативными запросами" можно понимать список слов и словосочетаний (запросов), который определенным образом связан с исходным запросом и формируется автоматически.

В поисковой машине компании Рамблер такой список выдается на ответной странице вместе с результатами поиска. В режиме просмотра найденных документов, принятом по умолчанию, ассоциированные запросы расположены в нижней части ответной страницы, после всех найденных документов. Они снабжены заголовком "***У нас также ищут...***". Это сокращенный список запросов, тематически связанных с исходным. Пользователи, которые часто обращаются к ассоциациям, могут перейти к другому режиму просмотра найденного, в котором доступен полный список ассоциаций к исходному запросу, и его удобно просматривать, т.к. он находится в самостоятельной зоне слева.

Таким образом, каждый посетитель сайта поисковой системы имеет возможность сравнить свой исходный запрос с теми формулировками и вариантами запросов, которые использовали другие пользователи, и, при желании, выбрать вариант, максимально уточняющий тематику и смысл запроса. Например, пользователь набирает запрос ***nokia*** и получает список сайтов с описаниями моделей телефонов, предложения купли/продажи, адресами сайтов, где можно скачать инструкции к телефону, картинки и мелодии для телефона, информацию о российском и международном представительствах компании и т.п. Одновременно он видит список связанных или ассоциативных запросов, т.е. имеет возможность выбора и мгновенного перехода к любому элементу ассоциативного ряда: ***мобильные телефоны***, ***nokia мелодии***, ***nokia 3310***, ***siemens***, ***sms***, ***сотовик*** и т. д.

Еще примеры:

- исходный запрос: отдых на Кипре; список ассоциаций: отели на Кипре, погода на Кипре, аппартаменты, детский отдых на Кипре, карта Кипра, лимассол и т. д;
- исходный запрос: Чечня; список ассоциаций: Кавказ центр, Басаев, Чечня Кавказ, Ичкерия, война и т.д.
- исходный запрос: рефераты; список ассоциаций: банк рефератов, украинские рефераты, доклад, курсовая, диплом, рефераты по истории и т.д.

3. Как формируются ассоциации?

В общем виде, алгоритм автоматического формирования ассоциаций построен на следующем предположении:

запросы, поданные одним и тем же пользователем в течение некоторого промежутка времени (одной поисковой сессии) с большой вероятностью имеют одну и ту же тематику.

Когда пользователь, дав запрос ***x***, сразу после этого дает еще и запрос ***y***, можно предположить, что запросы ***x*** и ***y*** между собой связаны.

Если это предположение подтверждается в результате анализа запросов от других пользователей, т.е. мы видим, что не один, а несколько посетителей сайта после запроса ***x*** переходят к запросу ***y***, то всем новым пользователям, подавшим запрос ***x***, можно порекомендовать поискать еще и запрос ***y***. И наоборот, поискавшим запрос ***y*** предлагается попробовать еще и запрос ***x***.

Вероятность тематического совпадения запросов ***x*** и ***y*** увеличивается, если они заданы подряд, поэтому при автоматическом формировании ассоциаций целесообразно учитывать именно пары "соседних" или непосредственно следующих друг за другом запросов. Что касается промежутка времени, то можно принять за единицу измерения одну отдельно взятую поисковую сессию, но если она слишком велика, то имеет смысл ограничить временной промежуток формирования ассоциаций. Этот промежуток подбирается экспериментально и редко составляет более 1,5 часов.

Для формирования ассоциаций используются протоколы работы web-сервера поисковой машины. В протоколах есть данные о времени обращения, сетевом адресе пользователя (IP), уникальном идентификаторе пользователя и, конечно же, информация о самом запросе к поисковой машине. Уникальный идентификатор пользователя – строка, которая хранится на компьютере этого пользователя в области так называемых cookies. Строка формируется в момент, когда пользователь впервые попадает на любую web-страницу Рамблера и с этого момента меняется очень редко.

Автоматическая процедура формирования ассоциаций включает в себя несколько этапов:

1. построение списков "кандидатов в ассоциации";
2. объединение списков и формирование собственно ассоциаций;
3. ранжирование ассоциаций
4. фильтрация ассоциаций;
5. сохранение окончательного списка ассоциаций.

На первом этапе построения ассоциаций протоколы работы web-сервера обрабатываются специальной программой, которая строит списки запросов, поданные каждым пользователем. Для каждого запроса запоминается время, в которое он был подан и количество найденной информации. Запросы, поданные одним и тем же пользователем с интервалом не более 1.5 часа, объединяются в группы. Все пары запросов из таких групп считаются кандидатами на ассоциации.

Затем выполняется объединение групп и подсчет частот употребления пар запросов. При этом пара запросов, поданная одним пользователем, учитывается только один раз (не важно, как часто этот пользователь подавал эти запросы).

Программа группировки получает на входе множество троек вида: $(\boldsymbol{x}, \boldsymbol{y}, f_{xy})$, где ***x*** и ***y*** – запросы, а f_{xy} – количество пользователей, подавших данную пару запросов (частота совместной встречаемости). В процессе группировки для каждого запроса *x*, все тройки вида $(\boldsymbol{x}, \boldsymbol{y_1}, f_{xy1})$, $(\boldsymbol{x}, \boldsymbol{y_2}, f_{xy2})$, ... $(\boldsymbol{x}, \boldsymbol{y_n}, f_{xyn})$, $(\boldsymbol{z_1}, \boldsymbol{x}, f_{z1x})$, $(\boldsymbol{z_2}, \boldsymbol{x}, f_{z2x})$, … $(\boldsymbol{z_m}, \boldsymbol{x}, f_{znx})$ превращаются в список ассоциаций запроса ***x***: $\mathbf{assoc}(\boldsymbol{x}) = \{(\boldsymbol{y_i}, f_i)\}$. Для каждой ассоциации сохраняется ее частота.

Полученные списки ассоциаций вполне можно использовать для социологических, лингвистических и пр. целевых исследований аудитории Интернета. Однако, они непригодны для показа пользователям поисковой машины. Дело в том, что очень большое количество пользователей практически в каждой поисковой сессии дает запросы ***sex***, ***porno***, ***рефераты***, ***Москва***, ***знакомства*** и т.п. Получается, что в списке ассоциаций практически любого запроса в большом количестве присутствуют эти слова, и их частотность и повторяемость достаточно высока.

Возникает необходимость в специальном ранжировании ассоциаций для подавления избыточных ассоциаций. С этой целью для каждого запроса ***x***, имеющего список из *n* ассоциаций $\{(\boldsymbol{y_i}, f_i)\}$, i=1…n, где $\boldsymbol{y_i}$ – ассоциативный запрос, f_i – количество пользователей, которые искали одновременно ***x*** и $\boldsymbol{y_i}$ в течение 1.5 часов, вычисляются ранги элементов списка ассоциаций, и элементы сортируются по убыванию веса. Ранги определяются по следующей формуле:

$$rank(\boldsymbol{x}, \boldsymbol{y_i}, f_i) = S(\mathbf{assoc}(\boldsymbol{x}), \mathbf{assoc}(\boldsymbol{y_i}))^{\alpha} \cdot \left(\frac{f_i}{\max_{i=1\ldots n} f_i} \right)^{\beta} \cdot W(\boldsymbol{y_i}) \cdot Z(\boldsymbol{x}, \boldsymbol{y_i}) \quad (1),$$

где:

- $S(\boldsymbol{x}, \boldsymbol{y_i})$ – мера похожести списков ассоциаций запроса ***x*** и списка ассоциаций запроса $\boldsymbol{y_i}$ (косинусообразный коэффициент корреляции);
- $W(\boldsymbol{y_i})$ – функция от числа слов в запросе y_i, дающая небольшой приоритет запросам из двух или трех слов и уменьшающая ранг длинных запросов;

- $Z(\boldsymbol{x}, \boldsymbol{y_i})$ – функция, повышающая вес ассоциациям, текст которых целиком включает в себя запрос x;
- α, β - постоянные коэффициенты.

Мера похожести между списками ассоциаций $\{\boldsymbol{a_i}, f_i\}$ и $\{\boldsymbol{b_j}, f_i\}$ определяется следующим образом:

$$S\left(\{\boldsymbol{a_i}, f_i\}, \{\boldsymbol{b_i}, f_j\}\right) = \frac{\sum_{i,j,\ \boldsymbol{a_i}=b_j} f_i \cdot f_j}{\sqrt{\sum_i \left(f_i^2\right) \cdot \sum_j \left(f_j^2\right)}} \qquad (2)$$

Благодаря учету степени похожести в каждом списке ассоциаций вверху оказываются те запросы, в списке ассоциаций которых присутствует много запросов анализируемого списка. Это означает, что слова ***реферат***, ***Москва*** и др. получают высокий вес только в списках ассоциаций подобных слов.

И последний этап формирования ассоциаций - фильтрация запросов, включающих в себя ненормативную лексику.

4. Что ассоциации дают пользователям?

4.1. Уточнение или расширение запроса

Наиболее распространенная сфера применения ассоциативных запросов – это уточнение и детализация исходного запроса с целью повышения релевантности найденных документов. В результате процедура поиска информации существенно упрощается и занимает меньше времени.

Кроме того, простой просмотр списков ассоциаций позволяет выявить недостатки исходной формулировки запроса: его неоднозначность, возможность двоякого истолкования, "размытость". В результате посетитель поискового сайта имеет возможность на примере других запросов от других пользователей обучиться тому, как правильно задавать вопросы поисковой системе, т.е. по сути – воспользоваться "коллективным разумом".

Например, запрос ***театры*** приведет к появлению огромного количества найденных документов, в котором первые несколько тысяч будут вполне релевантными. Но театров много, и найти в этом потоке сайтов нужный – это непростая задача. Неопытный пользователь часто теряется при виде такого обилия информации. А просмотр списка ассоциаций, - ***театральная афиша, театры москвы, репертуар, малый театр, ленком, театр эстрады, кинотеатры, детские театры, репертуар театров***, - позволяет сделать вывод о необходимости сузить тематику запроса, т.е. указать, какой конкретно нужен театр, запросить афишу или билетную кассу и т.п.

4.2. Получение новых знаний по теме запроса

Часто бывает так, что пользователь впервые заинтересовался какой-либо темой. Например, он хочет найти информацию о каком-либо товаре, но пока не знает, кто выпускает данный товар, каковы преимущества той или иной фирмы и т.п. Список ассоциаций чаще всего содержит не только уточненные и переформулированные запросы, но и названия фирм, людей, событий, предметов, наиболее популярных и распространенных в данной тематике. Например, запрос ***велосипеды*** приводит к появлению ассоциаций: ***вело, детские велосипеды, горные велосипеды, продажа велосипедов, merida, велозона, веломир*** и т.д.

Видно, что автоматически сформированный список содержит не только наименования различных типов велосипедов (***горные велосипеды, детские велосипеды***), но и названия популярных торговых марок, и магазинов (***велозона, веломир***), брендов (***merida***) и так далее.

4.3. Исправление ошибок в написании запроса

В некоторых случаях ассоциации предоставляют пользователю даже такую возможность, как исправление ошибок в написании запроса. В принципе, ошибка при написании запроса чаще всего приводит к отказу поиска или к отсутствию релевантных документов в списке найденного. Например, типичная графическая ошибка – использование латинской раскладки клавиатуры для написания русского слова и наоборот. Поскольку эта ошибка является типичной, у нее есть шансы найти отражение в списке ассоциаций. Например, запрос ***gjujlf*** ("погода", набранная в латинской раскладке клавиатуры). Список ассоциаций: ***прогноз погоды, погода в москве, погода на неделю, weather*** и т.д.

Другой пример – обычные орфографические ошибки. Самая сложная ситуация для пользователя возникает, когда он не в состоянии эту ошибку найти. Обычно в таких случаях поиск оказывается неудачным. Но некоторые

наиболее распространенные ошибки отражены в ассоциациях, и это может помочь пользователю. Например, запрос ***фотоаПарат*** ассоциируется у поисковой машины со словом ***фотоаППарат***, что будет отражено в списке связанных запросов; или же запрос ***ТольяТи***, который приводит к появлению ассоциации ***ТольяТТи*** (правильное написание названия города).

5. Что ассоциации дают web-мастерам и разработчикам поисковых систем?

Ассоциации - это источник полезных данных о том, чем интересуются люди, посещающие те или иные сайты. Например, из списка ассоциаций по слову ***Nokia*** следует, что обладатели мобильных телефонов этой фирмы интересуются возможностью программирования мелодий. Поэтому для того, чтобы сервер, посвященный мобильным телефонам, был интересен и популярен, на нем надо разместить информацию по мелодиям для сотовых телефонов.

6. Что ассоциации дают рекламодателям?

6.1. Планирование рекламных кампаний и таргетинг

Большинство поисковых машин позволяют показать рекламные модули под конкретные запросы или блоки запросов. Преимущество таких показов очевидно:

- сужение круга пользователей, которым будет показана реклама, т.е., по сути, уменьшение стоимости рекламы,
- и при этом увеличение количества пользователей, потенциально заинтересованных в рекламе именно этой тематики, т.е. повышение отдачи от рекламной кампании.

Рекламодатели заинтересованы в развитии подобной тематически ориентированной рекламы, но нередко затрудняются составить исчерпывающий список потенциально интересных / выгодных запросов. Например, для сотрудника банка естественно включить в список запросов для рекламы своего банка слова ***виза*** и ***визы***. При этом подразумевается, что пользователи, задавшие эти запросы ищут пластиковые карты международной банковской системы VISA. В реальности же эти запросы неоднозначны, и могут относится как к банковской тематике, так и к туризму и путешествиям (виза для поездки, туристическая виза, и т.п.).

Ассоциации информируют пользователя о неоднозначности и помогут составить правильный список слов / запросов для показа рекламы.

6.2. Определение целевой аудитории

Некоторые запросы помогают понять, кто является пользователем услуги или товара, т.е. определить целевые характеристики аудитории (социальные, возрастные, сфера интересов и т.п.).

Например, аудитория канала СТС (пользователи, интересующиеся сайтом и чатом СТС) часто спрашивают ***земфира***, ***мату***, ***руки вверх***, ***ддт***, ***britney spears*** и **дельфин**, а вот у аудитории MTV другие музыкальные пристрастия: ***децл***, ***eminem***, ***limp bizkit***, ***prodigy***, ***сектор газа***.

Те, кто спрашивает ***parker***, интересуются также другими производителями авторучек, наручными часами, канцтоварами и кожгалантереей.

Конечно, информации, которая выдается на ответной странице поисковой машины для полноценных маркетинговых исследований недостаточно – для этого нужно использовать полные списки запросов, рассчитывать отношение частот ассоциаций к общим спискам запросов (affinity index) и т. д. Вся эта информация доступна, и при необходимости может специально обрабатываться и использоваться.

7. Как еще можно использовать ассоциативные запросы?

Предложенная методика формирования ассоциаций может использоваться для автоматического уточнения поисковых запросов. Так, например, при поиске по запросу ***гороскоп*** можно повысить вес (показатель релевантности) тем документам, которые соответствуют еще и ассоциированным запросам (***гадание***, ***сонник*** и т. д.).

В данной статье перечислены далеко не все возможности применения ассоциативных запросов. Они гораздо более обширны. Очевидно, что метод автоматического формирования ассоциаций перспективен не только как удобный

пользовательский инструмент, но и как средство для дальнейшей разработки механизма поиска и совершенствования функциональных возможностей поисковых систем.