Алгоритм автоматического создания фильтров для сортировки входящей почты

Свердлов М.И.(Mike@globalinfactory.com), Андриенко П.В.

Санкт-Петербургский Государственный университет аэрокосмического приборостроения

Автоматизация обработки потока электронной корреспонденции рекламной компании является важным элементом технологии Friendly Advertising (Дружественная реклама). Такая автоматизация устраняет узкое место в применении самой технологии Friendly Advertising, а именно позволяет эффективно обрабатывать большие потоки ответной реакции потребителей – их писем. Как правило, при этом к большому количеству получателей приходит одно и то же рекламное сообщение и некоторые из получателей на него отвечают отправителю. Таким образом, ставится задача обработки ответных сообщений от получателей рекламы.

В настоящее время распространен подход, когда входящие ответные сообщения просто не обрабатываются, или обрабатываются простейшими фильтрами в целях удаления адресов людей, не желающих получать рекламу (общепринятая форма запроса на удаление из списка рассылки: «REMOVE» в поле Тема письма). Тем не менее, обработка хотя бы первичных откликов от получателей позволяет существенно повысить эффективность такой рекламы, а также позволяет снизить возможное раздражение получателей рекламы. Причиной, которая приводит к отказу от обработки откликов, является отсутствие эффективного алгоритма автоматизированной обработки. Обработка же вручную, как правило, экономически нерентабельна.

Однако, практика авторов работы в данной сфере и известный авторам опыт коллег показывает, что на любую рекламную кампанию, включающую рассылку сообщения с одинаковым текстом, возможно ограниченное количество категорий откликов, часто до 10. Лишь незначительный процент от общего количества откликов невозможно классифицировать в эти категории. Каждая категория может отличаться типом реакции системы или оператора. Например, такой реакцией может стать отправка стандартного ответа, передача на дальнейшую обработку менеджеру по продажам, простановка в систему планирования для дальнейших контактов, внесение в особый список и т.д.

Задача автоматического распознавания категорий входящих откликов решается применением фильтров по ключевым фразам, возможность создания которых присутствует сегодня практически в каждой программе-клиенте электронной почты. Однако, в связи с этим возникает проблема создания фильтров, способных распознавать категорию сообщения. Цели, которые ставятся при создании фильтров, следующие. Во-первых, с помощью созданных фильтров должен распознаваться достаточно высокий процент сообщений, во-вторых, погрешность распознавания не должна превышать допустимую, в соответствии с целями обработки сообщений.

Фильтры могут создаваться вручную, однако в связи с этим возникает следующий комплекс проблем: относительно невысокий процент распознавания и относительно высокая погрешность, либо значительное время на создание и тестирование фильтров. Фильтры необходимо протестировать на значительной выборке сообщений (как правило, 1000 и более).

В данной статье предлагается алгоритм автоматизированного создания фильтров. Заметим, что данный алгоритм работает для английского языка. Реализация его для других

языков с иной структурой может представлять сложность, иметь худшие результаты, либо требовать доработки алгоритма.

Рассмотрим вначале простую модель создания фильтров по выборке значительного размера. Оператор определяет категории, вручную сортирует достаточно большое число сообщений (например, 2000) по категориям, после чего передает выборку в таком виде программе.

Принцип работы алгоритма заключается в сравнении частоты появления той или иной ключевой фразы в каждой категории. Если в одной категории фраза встречается значительно чаще, чем в другой, при этом не менее какого-то количества раз, то можно считать, что оно является ключевой фразой.

Поясним подробнее. Укрупненно, порядок выделения ключевых фраз таков:

1) создание списка всех слов во всех сообщениях каждой категории;
2) определение индекса присутствия для всех выражений (упрощенно – частота появления слова в категории);
3) выявление и пометка кандидатов в ключевые выражения – выражений, для которых индекс встречаемости выше заданного;
4) для всех таких кандидатов: добавление в список выражений с размерностью на одно слово больше, слово берется из сообщения справа от того, на которое оканчивается текущее выражение, и присоединяется к кандидату;
5) повторение шагов 2-4 для всех новых появившихся выражений, пока не перестанут появляться новые кандидаты в ключевые выражения на шаге 3.

Цель части алгоритма выше – выделить все возможные выражения из любого количества слов, которые повторяются в категории не менее заданного числа раз, то есть проявляют собой некоторую повторяемость, и следовательно, могут являться кандидатами в ключевые выражения.

6) сравнение частоты встречаемости кандидатов в ключевые выражения вне и внутри категории: если их частное не больше определённого значения, то выражение считается ключевым.

Таким образом, алгоритм определения выражения как ключевого базируется на двух критериях: достаточная повторяемость в распознаваемой категории и достаточно большое различие в частоте появления внутри и вне распознаваемой категории. Позволим себе небольшое отступление: небезынтересно, эти принципы определения критериев распознавания категорий объектов, по-видимому, соответствуют двум принципам естественной эволюции – размножению и селекции.

Итак, результатом работы алгоритма является набор ключевых выражений, каждое из которых способно с обозначенной погрешностью распознать принадлежность или непринадлежность нового сообщения к какой-либо категории.

Рассмотрим подробнее каждый из пунктов алгоритма.

1. Получив выборку сообщений, разбитую на категории, программа создает список всех слов, встречающихся в каждом сообщении в каждой категории. Слово – последовательность символов, разделенная пробелами, и рядом других знаков («!?;:%()+» и т.д.). В списке для каждого слова также указывается адрес сообщения в базе, в котором встречается слово, и позиция внутри сообщения. Эти параметры понадобятся для быстрого доступа к данному слову в сообщении при дальнейшем создании выражений, состоящих из 2, 3 и большего количества слов.

2. Далее программа сортирует этот индекс слов по алфавиту и подсчитывает для каждого слова индекс встречаемости. Сортировка позволит более быстро подсчитать индекс

встречаемости. Индекс встречаемости – это параметр, являющийся чем-то средним между количеством сообщений, в которых встречается данное слово, и количеством всех вхождений данного слова в категорию. Ценность многократного появления слова в 1 сообщении ниже, чем появления слова во многих сообщениях по 1 разу. Чтобы учесть этот принцип снижающейся предельной полезности (в целях распознавания категории) каждого нового повтора слова и вводится индекс встречаемости слова.

На основании индекса встречаемости далее будет производиться отбраковка кандидатов в ключевые фразы из всего списка фраз.

Индекс встречаемости можно высчитать следующим образом. Все выделенные слова, содержащиеся во всех сообщениях категории, проходят по порядку. Если в одном сообщении какое-либо слово встречается 1 раз, то к индексу его встречаемости в выборке добавляется единица. Если 2 повтора слова, то к индексу встречаемости прибавляется 1+1/k (например, 1+1/3), если 3 повтора, то 1+1/k+1/(2k) (в примере – 1+1/3+1/6), если 4, то 1+1/k+1/(2k)+1/(3k) (т.е. 1+1/3+1/6+1/9) и т.д. Общая формула очередного добавляемого слагаемого: 1/((N-1)k), где N – номер очередного повтора слова в данном сообщении, k – коэффициент значимости повторений, который может принимать значения от 1 и выше. Увеличение индекса встречаемости для данного сообщения можно высчитать и проще, по известной формуле суммы конечной убывающей геометрической прогрессии, где аргументом будет количество вхождений фразы в сообщение.

Поясним на примерах, как действует индекс встречаемости. Если слово в сообщении встречается 1 раз, то к индексу добавляется единица. Если 10 раз, то не 10 единиц, а лишь 2 или 3 единицы, в зависимости от выставленного коэффициента k. Таким образом, если человек ответил сообщением, где 100 раз повторяется одно и то же слово REMOVE, то значимость этого слова с точки зрения распознавания категории окажется лишь, скажем, в 3,5 раза выше, чем, если бы человек сказал REMOVE в своем сообщении всего 1 раз. Индекс встречаемости слова в категории, встречающегося 10 раз в 1 сообщении из категории – лишь 3, в то время как по 1 разу в 10 сообщениях – 10.

На взгляд авторов, целесообразно устанавливать k консервативно: таким образом, чтобы при больших значениях повторов в сообщении, индекс встречаемости оказывался близким к 2. k может равняться 5,2, тогда, при 100 повторах прирост индекса встречаемости, примерно равен 2, а при неограниченно большом числе повторов – не уходит далеко от 2,5.

Заметим, что выбор формулы для расчета коэффициента встречаемости не является чем-то, безусловно детерминированным, может быть выбрана и другая функция. Однако выбранная функция, по мнению авторов, неплохо реализует требуемое от нее поведение.

Итак, ко всем словам, выделенным из всех сообщений в каждой категории, рассчитан индекс встречаемости, демонстрирующий, насколько часто в данной категории появляется данное слово.

3. Далее определяются и маркируются те выражения, которые являются кандидатами в ключевые выражения. Признаком является индекс встречаемости не ниже порогового значения A. «A» может принимать значения от 1 и более. Лишь в особых случаях использования алгоритма может иметь смысл устанавливать значение A=1 (то есть появление выражения в выборке как такового хотя бы 1 раз). Обычно достаточно надежным признаком повторяемости, а значит, неслучайности выражения, может являться значение A, равное 3 или 4.

4. Следующим шагом, для кандидатов в ключевые выражения необходимо построить и добавить в список выражения с числом слов, на единицу большим. Для этого программа проходит по списку кандидатов в ключевые выражения, обращается по указанной в них ссылке на сообщение и позицию в нем, и берет следующее за текущим выражением слово. Это слово присоединяется справа к текущему выражению, после чего выражение добавляется в список.

5. Для новообразованных выражений алгоритм определяет индексы встречаемости, после чего маркируются кандидаты в ключевые выражения, и процесс добавления новых выражений с размерностью на 1 слово больше повторяется. Происходит это до тех пор, пока не окажется, что больше нет новых выражений, удовлетворяющих критерию кандидата в ключевые выражения.

6. После того как для каждой категории выделены кандидаты в ключевые выражения, определяется, какие из них удовлетворяют второму требованию к ключевому выражению: достаточное отличие встречаемости внутри и вне категории.

Для этого подсчитывается отношение индекса встречаемости к общему числу сообщений внутри и вне категории. Чтобы подсчитать индекс встречаемости вне категории, индексы встречаемости данного выражения по всем остальным кроме исследуемой категориям суммируются, а из общего числа сообщений вычитается число сообщений в данной категории, и сумма индексов встречаемости делится на число сообщений вне категории. Если оказалось, что нужное ключевое выражение отсутствует в списке другой категории, то его индекс встречаемости оперативно рассчитывается: по списку берется первое слово в выражении и исследуются все случаи его появления в той категории.

Итак, мы разделили вычисленную частоту встречаемости вне категории на частоту встречаемости выражения-кандидата внутри категории. Если отношение будет не выше «B», в этом случае выражение можно считать ключевым, и оно вносится в конечный список ключевых выражений. Параметр B может принимать значения от 0 и выше, обычно до 1. Значение выше 1 будет означать, что в рассматриваемой категории должно быть меньше вхождений выражения, чем вне.

Рассмотрим влияние исходных параметров и коэффициентов алгоритма на качество результатов его работы.

Исходными параметрами алгоритма являются: коэффициенты k, A и B; размер эталонной выборки; число и размер категорий.

Качество результатов работы алгоритма оценивается по следующим параметрам:

1) мощность сортировки – мат. ожидание доли сообщений, автоматически сортируемой на другой произвольно взятой выборке сообщений (из той же рекламной кампании) по созданным ключевым выражениям;
2) точность сортировки – мат. ожидание погрешности сортировки сообщения из той же рекламной кампании, но не входящего в эталонную выборку
3) могут рассматриваться 1) и 2) для каждого отдельного ключевого выражения и по совокупности функционирования всех выявленных ключевых выражений для каждой категории.

Чем выше коэффициент k, тем больше индекс встречаемости похож на значение числа сообщений, в которое входит выражение внутри категории, то есть меньше учитывается влияние повторяемости внутри сообщения. Если сделать этот коэффициент слишком строгим (высоким), то может пострадать как точность, так и мощность сортировки: могут быть упущены ключевые выражения, содержащиеся по несколько раз в сообщениях, причем содержащиеся уместно, отражая смысловую специфику категории. Если сделать коэффициент слишком мягким, то индекс встречаемости может завышаться в связи с нерелевантными повторениями ключевых выражений, например, как способа выражения эмоций, или вследствие многословия корреспондента, где он может, многократно возвращаясь к уже сказанному, объемно выражать одну и ту же мысль. В результате также и точность, и мощность сортировки будет страдать.

Чем выше коэффициент A, тем более строгие требования предъявляются к повторяемости выражения, чтобы определить его как кандидата в ключевые. Если сделать его слишком высоким, улучшается точность сортировки, но ухудшается мощность, если его

занизить, то наоборот. Таким образом, коэффициент А является параметром оптимизации. Параметр А влияет на количество ключевых выражений, которые лишь случайно в эталонной выборке соответствуют критериям выбора ключевых выражений и не отражают по смыслу критерии сортировки. Далее, при работе выбранных ключевых выражений на новых сообщениях, процент таких случайных ключевых выражений повлияет на точность сортировки. Формализация метода оптимизации на данный момент является для авторов неочевидной задачей, вероятно, она должна осуществляться путем эксперимента.

Чем выше коэффициент В, тем более строгие требования предъявляются к количеству отличия категории. Увеличение В снижает мощность сортировки, и увеличивает точность. Понятно, что чем больше эталонная выборка и более равномерно распределено количество сообщений по категориям внутри нее, тем более будет похож коэффициент В на мат. ожидание погрешности при дальнейшей сортировке. При небольших выборках (сотни сообщений) отличие мат. ожидания погрешности может отличаться от В в несколько раз, так что В необходимо задавать достаточно жестко, например, для 5% конечной погрешности и размере эталонной выборки 500 сообщений В может задаваться в размере 1%.

Количественные зависимости значений качества работы алгоритма от исходных параметров и вычисление требуемых и оптимальных значений параметров авторы оставляют специалистам по статистике.

Заметим, что в отсутствие статистически точных заранее предсказуемых значений погрешности, работа алгоритма требует выборочного контроля результатов для установления реальной степени погрешности. Если она выходит за пределы допустимого, необходимо пересмотреть параметры, влияющие на погрешность, и проработать эталонную выборку алгоритмом повторно. При недостаточной точности создания ключевых выражений алгоритмом автоматически, можно применить ручное модерирование результатов работы программы. Оператор при этом принимает решение о включении или невключении ключевого выражения, предложенного алгоритмом, в конечный набор ключевых выражений, на основании наличия смыслового соответствия между выражением и принципом сортировки в данную категорию.

Еще одним средством повышения точности сортировки является фильтрация одновременно всеми созданными фильтрами каждого сообщения. Если к сообщению одновременно успешно применяется более одного фильтра, сообщение считается спорным и оставляется для обработки оператором.

Рассмотрим некоторое усовершенствование алгоритма. Целью этого усовершенствования является снижение размера эталонной выборки без снижения качества работы алгоритма.

Вместо начальной большой, уже рассортированной на категории выборки используем начальную небольшую. Вышеописанный алгоритм проходит по выборке и выделяет ряд ключевых слов. С их помощью, алгоритм автоматически сортирует выборку большего размера, например в 3 раза. Далее алгоритм вторично применяется уже к новой выборке, рассортированной с помощью ключевых выражений, созданных им же на первом этапе. Благодаря тому, что в сообщениях внутри одной категории часто содержится одновременно более, чем одна ключевая фраза, оказывается возможным обнаружить на втором прогоне алгоритма новые ключевые выражения, отличающие категорию. Также этому способствует то, что на большего размера выборке по первому критерию ключевого выражения – повторяемости, проходит большее число ключевых выражений.

Далее, по выделенным на втором прогоне ключевым выражениям делается сортировка еще большей выборки, далее применяется алгоритм, и так до тех пор, пока не закончится весь объем сообщений, подлежащая сортировке на категории.

Возможной проблемой является накопление погрешности сортировки. Авторы не исследовали на практике поведение алгоритма при подобном способе использования, хотя не

исключают возможности автоматического «затачивания», то есть сохранения размера погрешности при увеличении количества сортируемых сообщений, при оптимальном подборе коэффициентов алгоритма. Соображения в пользу этого заключаются в том, что алгоритм включает механизм отсеивания случайных факторов, то есть ошибок, которые могут появиться вследствие недостаточности выборки. Если связи внутри сообщений (то есть наличие нескольких ключевых фраз, характеризующих данную категорию) и механизм отбраковки случайности окажутся сильнее, чем накопление ошибки, погрешность может и не накапливаться.

Отметим, что сфера наиболее эффективного применения алгоритма включает в основном массовые рассылки с весьма значительным числом получателей. Это обусловлено достаточно большим числом сообщений в эталонной выборке, необходимых для качественной работы алгоритма. Пример эффективного использования алгоритма: рассылка составляет 10 млн. сообщений, отклик 1% – 100 тыс. сообщений, ручная обработка 2000 сообщений, еще 50 тысяч сортируются автоматически с точностью 5%. Экономия по сравнению с сортировкой сообщений вручную составит порядка 4 рабочих месяцев (сортировка 500 сообщений за рабочий день), или $1600 по российским расценкам на рабочее время специалиста, либо порядка $10,000, по расценкам на рабочее время в США. Ручное составление и тестирование эффективных фильтров заняло бы около месяца, следовательно, экономия $400 по российским расценкам и $2500 по расценкам США.

Среди авторов нет специалистов по созданию алгоритмов, поэтому предполагается, что предложенный алгоритм допускает значительные оптимизации по времени на выполнение, а также требуемой памяти. Однако авторы оставляют его оптимизацию специалистам в области алгоритмизации, статистики и теории множеств. Задачей данной статьи было лишь показать принцип действия, позволяющий генерировать ключевые фразы для автоматической фильтрации сообщений по категориям.

Дальнейшее развитие алгоритма может включать механизм создания фильтров с регулярными выражениями (regular expressions). Использование регулярных выражений доступно в таком емейл-клиенте как TheBat!, а также в профессиональных системах обработки почты и управления отношениями с клиентами (CRM). Использование регулярных выражений, или хотя бы некоторых из них, таких как функция NEAR, могут позволить значительно увеличить мощность фильтров – долю автоматически обрабатываемых сообщений, по оценке авторов, до 80% всего количества сообщений. В то же время, здесь потребуются значительно более сложные механизмы тестирования, а также специальные меры для снижения погрешности. Погрешность классификации при использовании регулярных выражений в общем случае ощутимо выше, чем при использовании жестких ключевых фраз.

Также, дальнейшее развитие алгоритма может предполагать взаимодействие ручной и машинной обработки входящих сообщений и одновременное автоматическое накопление/обобщение опыта ручной сортировки. Каждое новое входящее сообщение пополняет опыт системы и участвует в классификации в дальнейшем. Если система не в состоянии автоматически определить по критериям категорию сообщения с заданной степенью достоверности, она передает его на обработку оператору. Действие оператора является основой для дальнейшей автоматической работы системы.

Еще одно направление развития алгоритма – создание механизма упрощенной коррекции набора ключевых фраз без необходимости повторной ручной сортировки большого начального количества сообщений в случае незначительной корректировки рекламного сообщения. В этом случае категории ответов могут остаться прежними, однако признаки распознавания отчасти измениться. Алгоритм будет должен на основании относительно небольшой выборки, отсортированной вручную, сделать выводы о характере изменений и соответственно предложить измененный набор ключевых фраз.

Данный алгоритм с определёнными модификациями, учитывающими специфику, вероятно, может использоваться не только для обработки отклика на рекламу прямыми почтовыми рассылками, но и для поддержки пользователей через электронную почту. Специфика заключается в существенно большем разнообразии входящих запросов и более частом изменении категорий и стандартных реакций.

Наконец, отметим, что данный алгоритм является первым шагом к технической реализации системы многоступенчатого рекламного диалога: Friendly Advertising («Дружественная реклама»), разрабатываемой одним из авторов в своей диссертационной работе.

ЛИТЕРАТУРА:

1. Burton Marcus, “Modern Marketing Management”, New York, Random House, 1980.
2. Hamel G., Prahalad C. K., “Competing for the future”, Boston, 1994.
3. Shapiro, Carl and Varian, Hal R. Information Rules: A Strategic Guide to the Network Economy Harvard Business School Press, 1998.
4. Винер Норберт, «Кибернетика и общество», изд-во «Тайдекс-Ко», 2002.
5. Почепцов Г. Г., «Коммуникативные технологии двадцатого века», Москва, «Рефл-Бук», 1999.

Томилов В.В. “Культура организации международных коммуникаций”, издательство Санкт-Петербургского Университета Экономики и Финансов, 1996.