Рубрикатор в борьбе со спамом

Козеренко Анастасия Дмитриевна
ИРЯ РАН, м.н.с.
ЗАО "Ашманов и партнеры", лингвист
akozerenko@mail.ru

В докладе рассматривается применение рубрикатора при обработке писем почтовым фильтром. Традиционно рубрикатор позволяет решить задачу определения тематики текста. Однако в случае, когда перед нами спамерское письмо, задача усложняется: требуется определить не только тематику текста, но и его принадлежность к спаму. В докладе рассмотрены методы, которые для этого используются.

Данный доклад посвящен проблемам фильтрации содержания в почтовом фильтре спама, выполненном ЗАО «Ашманов и партнеры» (ср. Ашманов, Власова, Зоркий, Иванов, Калинин 2003).

Под спамом понимается несанкционированная почтовая рассылка, преимущественно коммерческого или рекламного характера. В спамерском письме адресату, как правило, предлагают приобрести какой-либо товар или услугу, а иногда и просто перевести немного денег на указанный счет. Попадаются и бескорыстные сообщения – например, так называемые «письма счастья» или письма, содержащие случайный текст. Тематика спамерских писем достаточно широка – от предложения купить коврик с электроподогревом или посетить порносайт до пошаговой инструкции, как заработать деньги в Интернете.

Задача фильтрации спама в потоке поступающих писем решается разнообразными способами. Среди них есть формальные – например, по специфическому содержанию заголовков письма, принадлежности исходного адреса или почтового сервера «черным спискам», и лингвистические. Мы остановимся на фильтрации спама методом контент-анализа, т.е. на фильтрации содержания собственно текста письма.

Для решения задачи квалификации письма как относящегося или не относящегося к спаму применяется рубрикатор, позволяющий причислить письмо к той или иной категории. Категории рубрикатора соответствуют видам спамерских писем, т.е. определяют не столько общую тематику письма, сколько тематику в рамках «спамерского диапазона»; каждая категория содержит набор соответствующих ключевых слов и словосочетаний, наличие которых проверяется в тестируемом письме. Всего рубрикатор насчитывает порядка тридцати подобных категорий, ср.:

- Азартные игры
- Заработай много денег
- Средства для похудания
- Переводы
- Базы данных

- Массовые рассылки
- Бизнес сувениры с вашей символикой
- Курсы английского языка

Письмо, содержащее достаточное количество ключевых слов одной или нескольких категорий, признается спамом.

Однако не все «спамерские» категории в рамках рубрикатора равноправны, т.е. в одинаковой степени свидетельствуют о том, что данное письмо является спамом. Условно категории рубрикатора можно разбить на три группы:

- «спамерские»
- причисляемые к спамерским
- «не спамерские» (но тем не менее регулярно встречающиеся в письмах, являющихся спамом)

К первой группе относятся такие категории, как «Общие признаки спама», «Халява» и некоторые другие. Эти категории содержат ключевые слова, наличие которых в письме позволяет однозначно заключить, что перед нами именно спам, ср.:

- извините за несанкционированную рассылку!
- приносим свои извинения за доставку вам этой информации путем массовой рассылки
- данная рассылка является разовой и не навязывает платных услуг
- это сообщение не является спамом
- рекордно низкие цены!!!
- бесплатно!!!

Ко второй группе относятся категории «Для взрослых», «Знакомства», «Заработок в Интернете» и некоторые другие, ср. ключевые слова таких категорий:

- live web-cam for free
- Britney Spears naked
- find love online
- заработай деньги в сети!
- хотите дополнительно заработать?
- заработать деньги на своем домашнем компьютере не выходя из дома

Теоретически, письмо такой тематики не обязано являться спамом. Личное письмо также может содержать предложения знакомства, дополнительного заработка или упоминание скрытых видеокамер. Однако при наличии нескольких подобных ключевых слов в тексте письма оно автоматически причисляется к спаму. Вероятность ошибочного определения письма подобной тематики оказывается достаточно малой.

Наконец, к третьей группе относятся такие категории, как «Туризм», «Недвижимость», «Полиграфия», «Телефония» и другие, собирающие часто предлагаемые или рекламируемые товары и услуги. Основные сложности при фильтрации содержания возникают с ключевыми словами именно этих категорий. Проблема заключается в том, что если в первых двух случаях определяемая тематика письма сама по себе является спамерской (или причисляется к таковым), и тем самым определение тематики письма автоматически задает и его принадлежность спаму, то отнесение письма к такой тематике, как «Туризм» еще ничего не говорит о том, является письмо спамом или нет. Для того чтобы категория «Туризм» в нашем рубрикаторе определяла также и принадлежность текста спаму, на ключевые слова этой тематики необходимо наложить определенные ограничения.

Так, следующие ключевые слова и словосочетания определяют общую тематику текста «Туризм», но не позволяют сделать выводы о его принадлежности спаму:

- бронирование авиабилетов

- визы и загранпаспорта
- групповые туристические программы
- круиз на теплоходе
- обзорная экскурсия по городу

Рассмотрим основные компоненты, которыми должны обладать ключевые слова таких категорий, как «Туризм», для того, чтобы однозначно задавать принадлежность текста спаму:

- эксплицитно выраженная рекламная составляющая (указание на высокое качество, уникальность услуги или товара), ср.:

  лучшие туры на лето
  эксклюзивный автобусный тур
  специальные супертарифы
  самый большой выбор пансионатов
  уникальная возможность посетить
  спецпредложения на летний отдых
  незабываемый отдых для вас

- эксплицитно выраженное предложение приобрести товар или услугу, ср.:

  предлагаем однодневные городские и загородные экскурсии
  предлагаем вашему вниманию бизнес-туры
  приглашаем в детские летние лагеря
  приглашает вас на отдых и лечение
  добро пожаловать в мир отдыха!

- наличие квантора всеобщности:

  туры на любой вкус
  авиабилеты в любую точку мира
  любые авиабилеты на международные рейсы
  билеты 24 часа в сутки!

- указание на срочность:

  визы шенген срочно!
  визы во все страны срочно
  срочные предложения по турам
  горящие туры!

- призыв к контакту с последующей сделкой:

  по вопросам приобретения билетов обращайтесь по тел.
  решили отдохнуть – звоните!

- упоминание скидок, распродаж, низких цен (в контексте туризма):

  суперцены на туры
  vacation blowout
  авиабилеты по очень хорошим ценам
  суперснижение цен на заезд

- упоминание продажи и доставки, в особенности бесплатной (в контексте туризма):

  авиабилеты с бесплатной доставкой
  доставка авиа и ж/д билетов
  билеты – продажа, доставка

- использование в ключевых словосочетаниях таких характерных лексических единиц, как *лучший, самый, специальный, эксклюзивный, уникальный, незабываемый, горящий,* приставок *спец-, супер-* , словосочетаний *24 часа в сутки, 7 дней в неделю*, и т.п., ср. вышеперечисленные термины.

Ключевые слова, содержащие один или несколько из перечисленных компонентов, позволяют не только установить, что данное письмо посвящено теме туризма, но и однозначно отнести его к спаму.

Разумеется, в некоторых случаях и такая подборка ключевых слов может не спасти от ошибочной квалификации письма. Например, список туристических предложений, разосланный от лица туристической фирмы своим коллегам, с достаточно большой вероятностью будет определен фильтром как спам. Помочь избежать подобных ошибок могут формальные методы (например, включение соответствующих адресов в «белые списки»).

Литература

1. Ашманов И.С., Власова А.Е., Зоркий К.П., Иванов А.П., Калинин А.Л. Технология фильтрации содержания для Интернет // Труды Международного семинара Диалог'2002 по компьютерной лингвистике и ее приложениям. Том 2. Москва 2002