XXXII Неделя науки СПбГПУ. Материалы межвузовской научно-технической конференции. Ч.V : С.73-74

УДК 004.7

Е.И.Балакин (6 курс, каф. АиВТ), В.М.Ицыксон, к.т.н., доц.

РАЗРАБОТКА АДАПТИВНОЙ СИСТЕМЫ ФИЛЬТРАЦИИ ЭЛЕКТРОННОЙ ПОЧТЫ

Проблема массовых нежелательных коммерческих почтовых рассылок через систему электронной почты по праву считается одной из главных бед современной сети Интернет. Колоссальные объемы навязчивой и электронной рекламы по разным оценкам составляют до 50% общего объема пересылаемых электронных писем, создают дополнительный трафик, увеличивают нагрузку на почтовые сервера и приводят к заметным тратам времени пользователя e-mail на просеивание входящей корреспонденции.

Существующие системы фильтрации электронной почты не всегда удовлетворяют предъявляемым к ним требованиям уровня защиты от массовых рассылок по ряду признаков, таким как несовершенство системы детектирования нежелательной почты, отсутствие региональной адаптации (то есть, направленность на фильтрацию исключительно англоязычных рассылок), высокая стоимость коммерческих продуктов.

Методы фильтрации спама можно разделить на два основных класса – анализ письма по формальным признакам и по содержанию. К первому относятся такие методы, как фильтрация по почтовым адресам отправителя, IP-адресам почтовых серверов, наличию в письме определенных формальных признаков (наличие служебных полей, отсутствие отправителя, получателя и т.п.). Методы второго класса фактически являются методами лингвистического анализа и реализуют фильтрацию по содержанию письма (ключевые словосочетания, статистика слов) и распознавание по образцам писем. Наибольший эффект достигается при использовании в системе фильтрации интегральной оценки по результатам анализа с помощью методик из обеих вышеупомянутых групп.

Если фильтры по формальным признакам способны функционировать с однократно настроенной базой правил, то фильтрация по лингвистическим признакам, дающая наилучшие результаты, требует постоянного обновления в связи с изобретательностью создателей массовых рассылок. Как правило, пользователи и администраторы программ фильтрации не располагают достаточной квалификацией или временем для требуемой настройки, поэтому наиболее перспективной является разработка программной системы, обеспечивающей автоматическую адаптацию правил анализа под изменяющийся поток нежелательных писем. Применение чисто вычислительных методов представляется проблематичным, так как последней тенденцией стало возникновение нежелательных рекламных писем, учитывающих психовизуальные особенности восприятия текста человеком для маскировки истинного содержания письма, например контраст цветов текста и фона, дублирование букв и слогов, в связи с чем, требуется реализация системы фильтрации, использующей элементы систем искусственного интеллекта.

Разрабатываемая система должна проводить, по крайней мере, бинарную кластеризацию писем на основе результатов предварительного обучения на конечной выборке электронных писем. В качестве математического аппарата для проведения анализа рассматривается возможность использования самообучающихся искусственных нейронных сетей типа ART (теория адаптивного резонанса), а также статистического метода опорных векторов.

На этапе предварительной обработки электронного сообщения осуществляется выделение основной части письма, его приведение к единой кодировке и текстовому формату, предфильтрация, с последующим построением векторной модели документа для

его отображения в пространстве признаков. Для векторизации текста возможно использование *n*-грамм, либо одной из класса ядерных моделей.

При использовании модели *n*-грамм текст представляется в виде вектора частотного распределения символьных последовательностей из *n* символов. В этом случае содержание письма представляется вектором **V** размерности M^n, где M – мощность множества символов алфавита, v_i – частота появления i-той *n*-граммы в тексте. Данное представление основано на предположении об информационной зависимости частотного распределения *n*-грамм в тексте от его содержания и обеспечивает постоянную размерность вектора документа. В последствии данное представление документа передается на вход системы кластеризации, например, ассоциативной нейронной сети.

Ядерные модели предполагают вычисление оценочной функции сходства множества документов по цепочкам символов или слов. Данный способ основан на использовании ядерных функций (радиальных базисных функциях), осуществляющих проекцию данных в многомерное пространство признаков, в котором осуществляется двоичное разделение. В результате, можно получить оценку вхождения непоследовательных цепочек символов или слов в текст и вычислить степень их сходства.

С помощью комбинации вышеперечисленных методик представляется возможным реализация программной системы фильтрации, способной быть устойчивой к различным модификациям нежелательных почтовых рассылок.