УДК 519.004.054

58 MIPT Conference theses (Sobolev)

Реализация и сравнительный анализ алгоритмов автоматической классификации электронной почты пользователя

К.В. Соболев, А.А. Соболев

Московский физико-технический институт (государственный университет)

В наши дни применение электронной почты многогранно: от банального общения, до синхронизации бизнес-процессов, что делает её неотъемлемой частью жизни человека. По данным сайта thenextweb.com, каждую минуту в мире отправляется свыше 168 млн электронных писем, а за день эта цифра превышает 188 млрд. При этом количество зарегистрированных адресов составляет более 3,3 млрд. Следовательно, в среднем на каждый почтовый ящик приходит порядка 60 писем в день, при этом часть писем не представляет большой ценности для получателя. Для того, чтобы пользователь не тратил много времени на разбор писем существуют специальные системы классификации почтовых писем.

В данной работе будут реализованы и сравнены эффективности алгоритма Наивного Байеса и метода опорных векторов в решении задачи классификации спама. В ходе данной работы планируется расширить уже имеющийся проект почтового клиента Mail Project. Клиент использует open source framework libMailCore, в котором уже реализован основной функционал для получения и отправки почты. Вся почта будет сохраняться в базу данных CoreData. Mail Project так же использует наш модуль фильтрации писем EFNN, который в данный момент сортирует при помощи нейронной сети. Модуль будет дополнен новыми методами классификации. Фильтрация писем будет осуществляться фоновым процессом, который будет периодически активироваться и проверять базу данных на наличие новых писем и сортировать их, если есть не отсортированные. Основным критерием оценки эффективности алгоритмов будет являться точность классификации. Обучение алгоритмов будет производиться на наборе почтовых писем американской компании энергетической компании Enron, которая обанкротилась в 2001 году и выложившей в общий свою почтовую базу данных. Преимуществом такого набора данных является то, что он уже отсортирован на спам и обычные письма, что позволяет точно определить точность работы наших алгоритмов.

Дальнейшая работа заключается в расширении данных алгоритмов классификации на сортировку писем по многим темам.

Литература:

1. А. А. Соболев, А. В. Соболева - Применение нейронной сети для автоматической классификации электронной почты пользователя // Труды 57 научной конференции МФТИ, 2014 год
2. Cambridge University Press. - Text classification and Naive Baye http://nlp.stanford.edu/IR-book/pdf/13bayes.pdf
3. Cambridge University Press. - Support vector machines and machine learning on documents http://nlp.stanford.edu/IR-book/pdf/15svm.pdf