УДК 004.931

П.К.Семенов (5 курс, каф. ИУС), Т.В.Коликова, ст. преп.

АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ ТЕКСТОВ КАК СРЕДСТВО БОРЬБЫ СО «СПАМОМ»

С середины прошлого века в мире началось повсеместное наступление информационной эпохи. Это привело к возникновению новых потребностей и как следствие новых угроз. В особенности это проявилось во многих профессиональных сообществах. Так создание общего информационного пространства (глобальные сети Интернет) позволило повысить оперативность доступа к необходимым сведениям, но и повысило риск лишиться конфиденциальных сведений.

Информационная эпоха породила многочисленных злоумышленников. И реалии современного общества таковы, что необходимы новые средства борьбы и защиты.

Данная работа посвящена методам защиты от нежелательных рассылок – «спама». В чем заключается сложность проблемы? Формы «спам»-писем настолько многолики, что даже пользователям иногда не удается его распознать. А проблему решить хотелось бы при помощи машин и без участия человека. С какими видами несанкционированной рассылки мы встречаемся достаточно часто?

1) Реклама. Эта разновидность проблемы встречается чаще всего – некоторые компании рекламируют свою продукцию с помощью «спама». Они совершают массовые рассылки либо самостоятельно, либо при помощи организаций, специализирующихся в этой области.
2) Мошенничество. Иногда «спам» используется для того, чтобы выманить деньги/номера кредитных карт/пароли доступа у доверчивых пользователей.
3) Заражение. Вредоносные программы также могут распространяться с помощью электронной почты.

Какие существуют способы борьбы со «спамом»? Самый простой и самый надежный – не позволить злоумышленникам узнать твой электронный адрес. Это достаточно трудно, но возможно при соблюдении мер предосторожности при работе в сети Интернет. Можно использовать специальное программные системы для автоматического определения «спама» в общем потоке входящих писем. Это ПО принято называть фильтрами. О способах построения таких систем и будет рассказано в данной работе.

В рамках данной работы был разработан программный инструментарий, реализующий алгоритмы и методы автоматической классификации текстов. Однако эти алгоритмы являются общими и могут быть применены для решения также более формальных задач. Например, для классификации математических векторов. Что в принципе позволяет применять инструментарий для решения и других задач распознавания.

Сама по себе автоматическая классификация текстов является стыком следующих областей знаний:

1) Natural Language Processing – подобласть информатики, направленная на обработку документов, написанных на естественном языке. Например, морфологическая разметка, выделение основы слов, построение синтаксических деревьев.
2) Informational Retrieval – подобласть информатики, направленная на получение и поиск информации в неструктурированных данных. Например, индексация различных мультимедийных и текстовых документов.
3) Machine Learning – подобласть математической теории искусственного интеллекта, направленная на изучение самообучающихся алгоритмов. Применяется в различных задачах распознавания образов, например, в автоматической классификации векторов.

Разработанная программная модель стала базой для оценки качества работы классификаторов. При моделировании использовались выборки из следующих данных в качестве обучающих и тестовых коллекций:

1) так называемая Bruce spam-коллекция с web-ресурса http://untroubled.org/spam/;
2) архив новостей UseNet'а в качестве коллекции не полезных сообщений.

Приведем некоторые полученные результаты (табл. 1, 2), но заметим, что они были достигнуты на приведенных выше данных и поэтому в какой-то мере обусловлены их характеристиками. При расчете под ошибками понимались как ошибка первого рода, так и ошибка второго рода без какого-либо различия по их возможной значимости.

Таблица 1. Алгоритм классификации текстов на основе морфологической разметки.

обучающая коллекция (шт.)		тестовая коллекция (шт.)		время обучения (с)	% ошибок
не "спам"	"спам"	не "спам"	"спам"		
3534	3534	435	435	84,95	0
2234	2234	435	435	43,14	0
1334	1334	435	435	22,9	0
682	682	435	435	11,64	0,12
890	890	227	227	14,67	0
458	458	227	227	8,09	0
227	227	227	227	3,75	0
227	227	51	51	3,76	0

Таблица 2. Алгоритм классификации текстов на основе алгоритма Пауля Грехэма (с порогом 0.9).

обучающая коллекция (шт.)		тестовая коллекция (шт.)		время обучения (с)	% ошибок
не "спам"	"спам"	не "спам"	"спам"		
3534	3534	435	435	116,50	26,67
2234	2234	435	435	70,37	31,26
1334	1334	435	435	37,01	26,67
682	682	435	435	19,37	27,82
890	890	227	227	23,95	28,85
458	458	227	227	13,21	24,67
227	227	227	227	5,98	25,33
227	227	51	51	6,09	23,53