МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ» (НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)
Факультет информационных технологий
Кафедра общей информатики

Разработка алгоритма детектирования ботов в социальных сетях

Выпускная квалификационная работа бакалавра

Соискатель: Зайков Д.И., гр.1208

Научный руководитель: Аникин Ю.А., к.т.н., н.с. ИТ СО РАН

Новосибирск 2014

Актуальность

Наиболее опасный для доверия в интернете вид киберпреступности – социальные боты.

Социальный бот - програма, имитирующая действия человека в социальной сети.
Объем рынка: от $40 млн. до $360 млн (Micheli, Stroppa “Twitter and the underground market”, 2013)

Типы ботов: рекламные боты, новостные боты, боты для раскрутки аккаунта, твита, тега, боты приложений, “аватары”, люди-“ретвитеры”.

Постановка задачи

Цель: Разработать алгоритм детектирования рекламных ботов для задачи фильтрации трафика социальных сетей.

Социальная сеть - Twitter.

На входе: профили и твиты

На выходе: классифицированные профили

Обзор литературы

Задача детектирования бота – задача классификации, похожа на задачу фильтрации спама в почте. L. Zhuang, etc. "Characterizing Botnets from Email Spam Records"

Алгоритм Байеса - наиболее оптимален для классификации текста. Alex Hai Wang “Detecting Spam bots in Online Social Networking Sites”

Модульная структура , где модуль принятия решений осуществляет конечную классификацию. Zi Chu, Steven Gianveccio “Detecting Automation of Twitter Accounts”

Деревья решений оптимальны для модуля принятия решений Zi Chu, Steven Gianveccio “Detecting Automation of Twitter Accounts”

Honeypot - ресурс-приманка, цель которого состоит в том, чтобы подвергнуться нападению злоумышленника. Securitylab.ru SecurityLab Технология Honeypot [Электронный ресурс]

Исследование

Наивный алгоритм Байеса – для классификации сообщений пользователя на “спам” и “не спам” по тексту сообщений пользователя Twitter (“твитов”, 140 символов).

Деревья решений - для классификации по атрибутам профилей пользователей.

Выбор атрибутов для классификации:

Спам-статус – число, дающее суммарную вероятность спама в 10 сообщениях.

Репутация – соотношение подписчиков и читаемых.

Description – текст в разделе профиля «о себе».

Реализация.Выбор инструментов

Язык программирования **Java**.

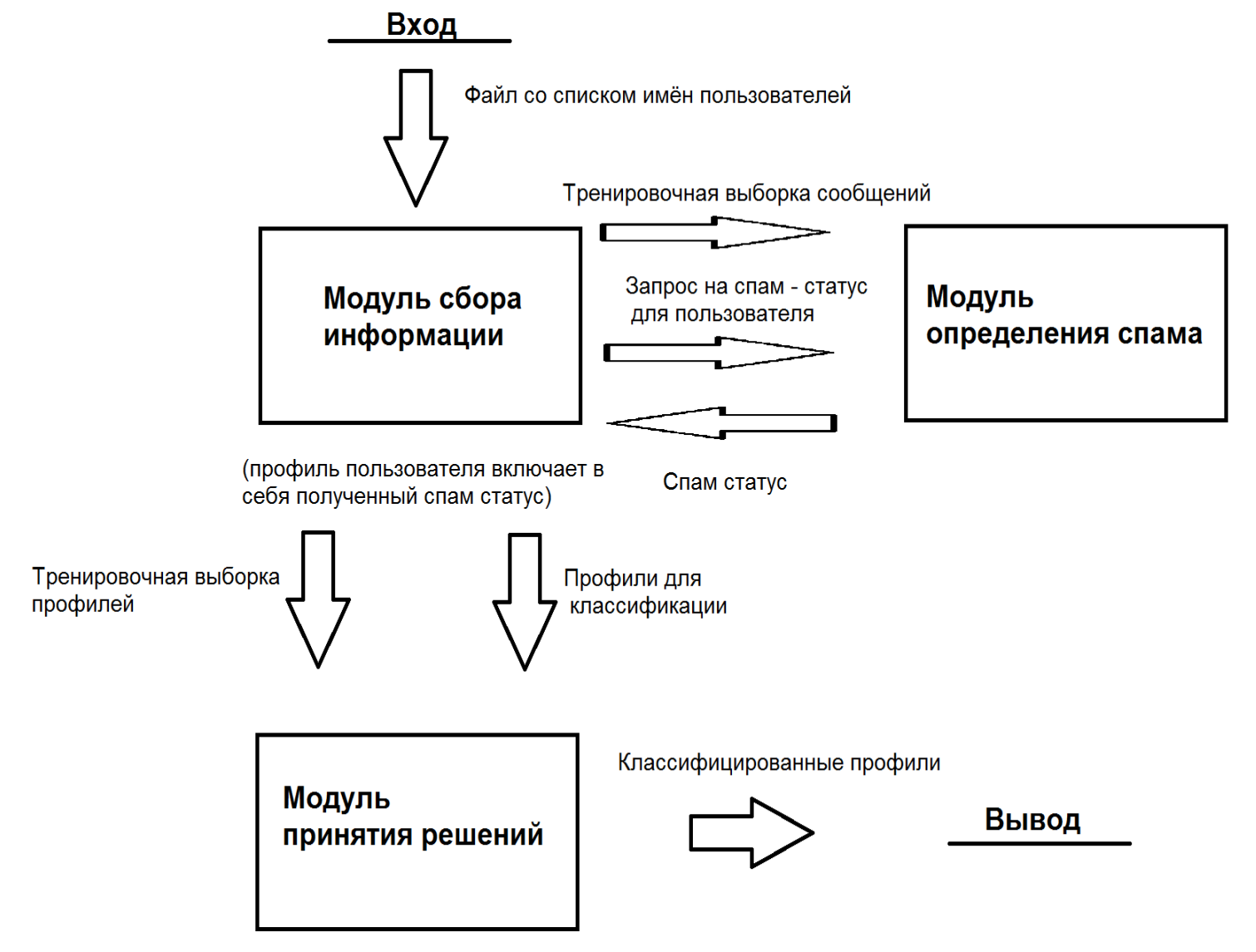
Twitter4j – неофициальная, свободно распространяемая библиотека для удобного использования Twitter API в собственных программах на Java. Работает с Twitter API посредством протокола OAuth. В проекте используется версия 3.0.5.

- REST-протокол для получения информации о твитах, пользователях, читателях и читаемых
- Сбор твитов в режиме Streaming

Weka – открытая библиотека алгоритмов машинного обучения для решения задач интеллектуального анализа данных. Реализована и API на Java. Есть GUI.

- Десятки классификаторов, в т.ч. деревья и таблицы решений, кластеризация, наивный Байес и др.

Реализация. Архитектура



Реализация. Модуль сбора информации

Входные данные – текстовый файл с именами пользователей

Выход: Данные в формате Arff - для работы и обучения классификатора

Модуль использует библиотеку Twitter4j для взаимодействия с Twitter API. Собрана информация из 500 профилей пользователей, 2500 сообщений для обучения и тестирования модуля спама и модуля принятия решений

Пример файла формата arff:

```
'Дистанция - душа красоты.',0.1,0.000,bot
'none',0.7,0.047,bot
'#здоровье Занимаюсь йогой, веду здоровый обрах жизни',0.05,0.227,bot
'#зарабатывайобщаясь Делаю прогон по твиттер аккаунтам. Цены приятные',0.0,0.218,bot
'none',0.75,1.000,bot
'Архитектор  # Картины из соломки - Походы на природу с палатками .',0.2,0.474,bot
'Вертебролог ...Искусственные цветы ,',0.0,0.103,bot
'Бизнес и коммуникации',0.4,1.618,bot
```

Реализация. Модуль определения спама

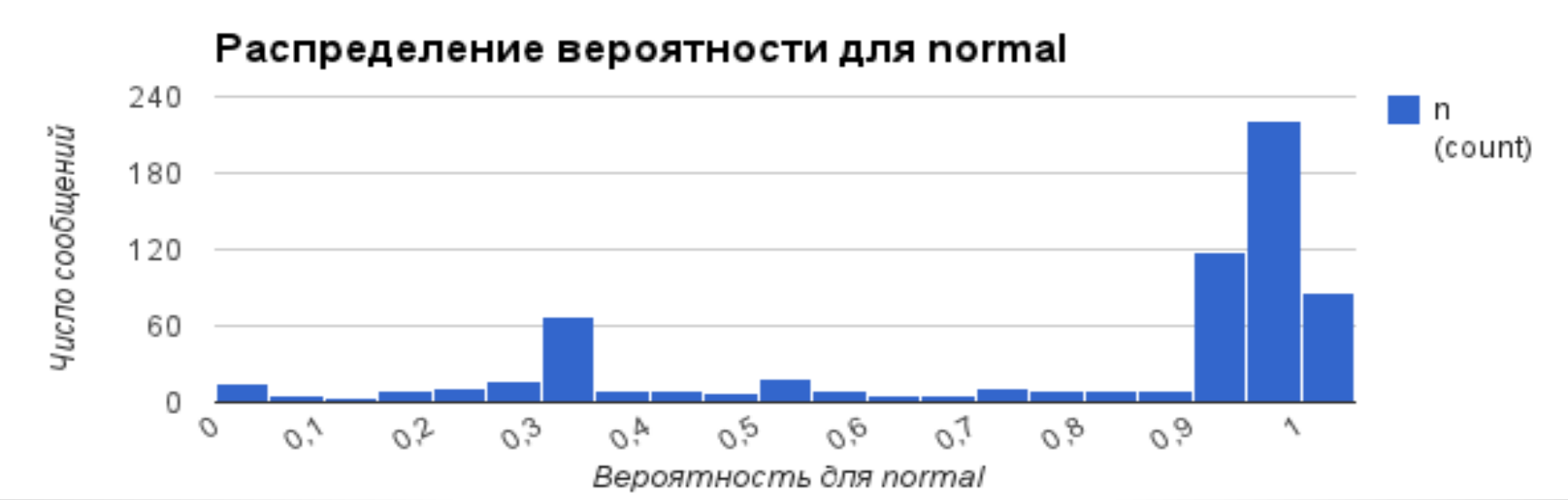
		Классифицировано, сообщений			
		spam	normal	Всего	Верно,%
На самом деле, сообщений	spam	395	218	613	64%
	normal	137	483	620	78%

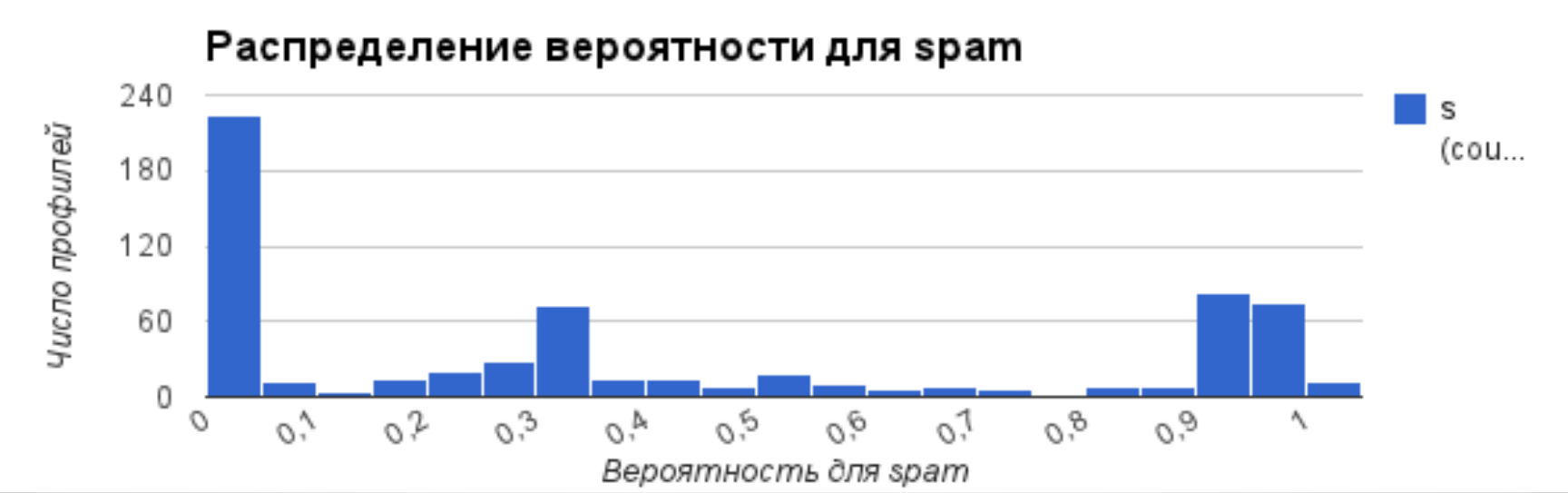
$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Точность(Precision)	Полнота(recall)
71% (0.716)	71% (0.712)

Реализация. Модуль определения спама





Реализация. Модуль принятия решений

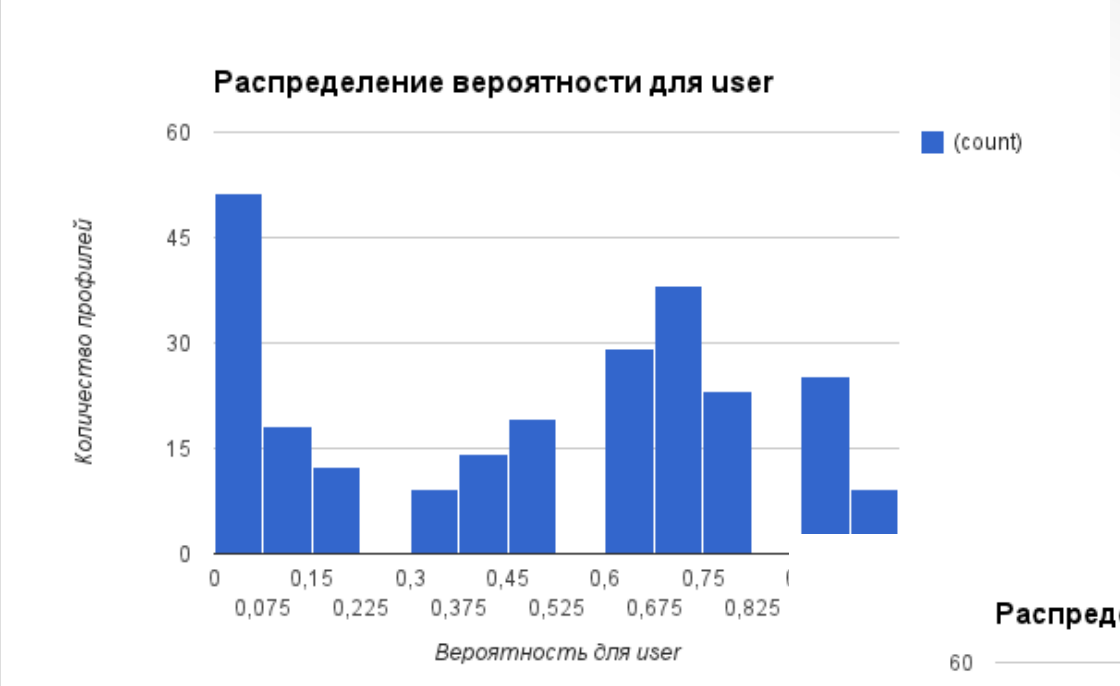
		Классифицировано, профилей			
		bot	user	Всего	Верно,%
На самом деле, профилей	bot	91	32	123	73%
	user	32	92	124	74%

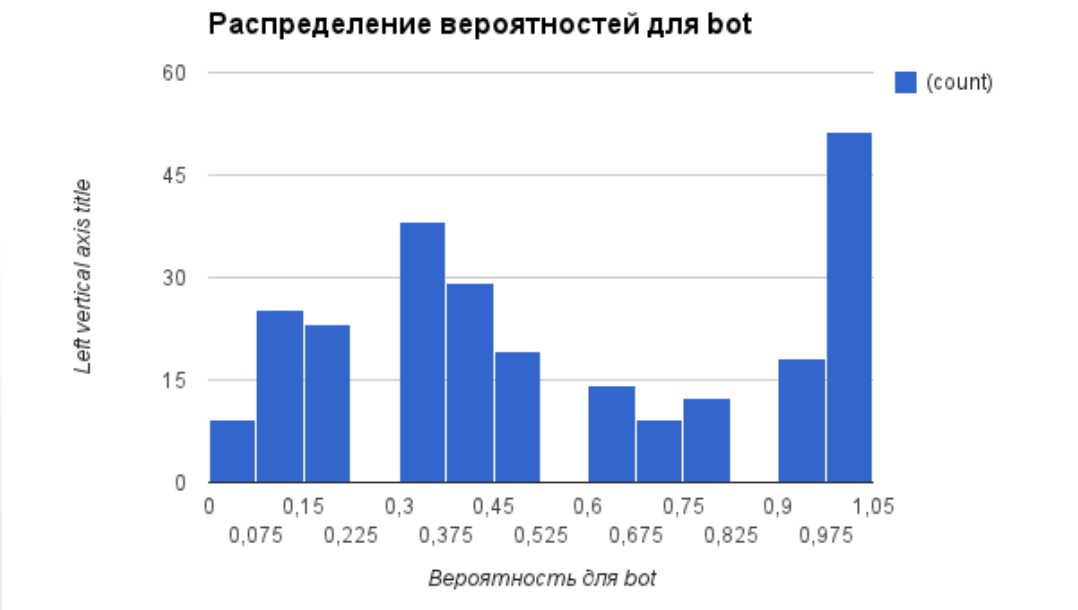
$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Точность(Precision)	Полнота(recall)
74% (0.741)	74% (0.741)

Реализация. Модуль принятия решений





Анализ результатов

Модуль определения спама классифицирует текст корректно с точностью 71%.

Модуль приятия решений классифицирует профиль корректно с точностью 74%.

Ограничения:

- Ворованные профили реальных пользователей,
- Боты “разбавляют” рекламные сообщения “человеческими”.Услуги по ретвиту сообщений реальных пользователей.
- Необходим анализ группового поведения пользователей в сети.

Перспективы развития:

Дополнительные модули: модуль оценки URL, модуль оценки периодичности и частоты сообщений(энтропии), модуль анализа группового поведения.

Заключение

Исследована предметная область. Бот-сети унаследовали значимость спам-сетей, активно развиваются и угрожают доверию в интернете.
Проведен анализ существующих методов. Подавляющее большинство алгоритмов основаны на классификации.
Разработан и реализован модуль сборки информации. Собрана тестовая и тренировочная информация.
Разработан и реализован модуль определения спама, определяющий спам-сообщения со средней точностью 71%.
Разработан и реализован модуль принятия решений, определяющий рекламных ботов со средней точностью 74%.
Результаты, полученные в работе, с учетом их развития, применимы во множестве задач анализа социальных сетей, например, в маркетинге и социологии. Методы можно использовать как для инструментов других систем, так и в самостоятельном программном сервисе.