ОПРЕДЕЛЕНИЕ ПОЛА АВТОРА КОРОТКОГО ЭЛЕКТРОННОГО СООБЩЕНИЯ

А. С. Романов (alexx.romanov@gmail.com)

Р. В. Мещеряков (mrv@keva.tusur.ru)

ГОУ ВПО «Томский государственный университет систем управления и радиоэлектроники», Томск, Россия

В статье рассматривается проблема определения пола автора короткого электронного сообщения длиной 20–200 символов. Приводится описание экспериментов и их результаты.

Ключевые слова: автор, пол, определение пола, электронное сообщение, короткое элекстронное сообщение.

GENDER IDENTIFICATION OF THE AUTHOR OF A SHORT MESSAGE

A. S. Romanov (alexx.romanov@gmail.com)

R. V. Meshcheriakov (mrv@keva.tusur.ru)

Tomsk State University of Control System and Radioelectronics, Tomsk, Russian Federation

Gender identifi cation of the author of a short message (20–200 characters) is studied. The paper describes a set of experiments with short message texts performed using a support vector machine approach. The task is viewed as a classifi cation problem with two possible alternatives: male and female. Important features of short messages to be considered when determining the author's gender are singled out. The database of electronic communications collected for research included 41780 posts by 15 men and 15women. Experiments used a software system Avtoroved developed by the paper's authors. Altogether, about 50 text attributes at the level of symbols, words, sentences and their combinations were studied. As a result, relevant characteristics of short messages were identified: unigrams and trigrams of symbols,function words, punctuation and emoticons. The total accuracy of gender identifi cations was 0.74.

Key words: author, gender, gender identification, message, short message.

Ежедневно миллионы людей общаются друг с другом посредством передачи коротких электронных текстовых сообщений с помощью системы SMS, электронной почты, интернет-пейджеров, социальных сетей и т.д. Среда и системы передачи сообщений становятся важной частью человеческой жизни и несут в себе важную информацию об интересах, привычках, социальном поведении людей. Мониторинг этой информации в определенные моменты времени и выявление лиц, имеющих целью совершение злонамеренных действий, становится актуальной практической задачей противостояния террористической угрозе и защиты государства. В беседе люди обычно выбирают манеру общения исходя из пола предполагаемого собеседника, поэтому создание интеллектуальной системы для определения потенциального злоумышленника в среде передачи сообщений и определение его психологического портрета целесообразно начать именно с определения пола автора сообщения. Известно, что сообщения мужчин проблемно ориентированы, кратки и содержательны одновременно. Женщины более общительны, их сообщения выразительны и эмоциональны [1].

Задача определения пола автора текста решалась многими зарубежными исследователями. Так в работе [2] для турецкого языка была получена точность правильного принятия решения о поле автора короткого сообщения из различных источников в сети Интернет близкая к 0,9 при использовании линейного дискриминантного анализа, а также лексических, морфологических, синтаксических характеристик текста. В работе [3] на корпусе англоязычных электронных писем с помощью метода опорных векторов и деревьев решений была достигнута точность идентификации пола 0,82 по функциональным словам и характеристикам уровня символов. Авторам работы [4] на примере британских эссе удалось достичь точности 0,81 путем анализа частоты встречаемости тетраграмм слов. Стоит отметить, что подобных исследований для русского языка не проводилось.

Проблема определения автора и пола автора короткого электронного сообщения имеет следующие важные отличия от других классических задач атрибуции текста, решавшихся авторами ранее [5]:

1. Небольшая длина сообщений (в среднем порядка 50–100 символов) по сравнению с другими типами текстов. Однако, как правило, существует возможность собрать большое количество сообщений для исследований.
2. Стиль сообщений одного автора от сообщения к сообщению может сильно меняться в зависимости от адресата: от формального в служебной переписке до неформального в частной.
3. Возраст, образование, сфера деятельности различных авторов-мужчин и авторов-женщин существенно варьируются. Формировать корпуса текстов для исследований следует с учетом этих особенностей. Также необходимо вводить корректирующие коэффициенты в итоговую методику.
4. Авторы могут умышленно скрывать информацию о себе или дезинформировать собеседников, выдавая себя за человека другого пола и гендера. Поэтому использование явных гендерно-окрашенных

признаков, таких как, например, глаголов в прошедшем времени в соответствующем роде, местоимений и т. д., не всегда представляется возможным. Таким образом, итоговый вектор признаков текста, описывающий соответствующий пол, должен состоять из слабоконтролируемых человеком характеристик, чтобы методика определения пола оставалась актуальной и для описанных случаев.

5. В коротких электронных сообщениях появляются дополнительные лингвистические элементы, такие как эмотиконы («смайлики»). Эмотиконы служат для придания написанным словам дополнительной эмоциональной окраски или для того, чтобы выразить эмоции по отношению, например, к предыдущей фразе собеседника. К дополнительным признакам также можно отнести использование Bulletin Board Code (BB-коды) — разметки текста, позволяющей расставить акценты в тексте путем выделении отдельных слов и фраз жирным шрифтом, курсивом и т. д.

Основной задачей, которую необходимо решить при определении пола автора, является получение репрезентативной и хорошо разделимой выборки числовых данных из корпуса текстов.

Формально проблема представляется как задача бинарной классификации произвольного сообщения $t \in T$ к одному из классов множества $C = \{C_1, C_2\}$, где C_1 — мужчины, C_2 — женщины. Целью является построение классификатора, решающего данную задачу, т. е. нахождение некоторой целевой функции $F : T \times C \to [-1,1]$, определяющей пол автора произвольного сообщения из множества T. При этом каждое сообщение рассматривается как вектор признаков $X = \{X_1, ..., X_n\}$. Обучение классификатора производится на сообщениях, пол авторов которых достоверно известен, т. е. существует множество пар $(t_i, c_j) \in D \subseteq T \times C$, где $t_i \in T$, $c_j \in C$.

Классификатор достаточно обучить один раз и более не переобучать, как это приходится делать при решении задачи идентификации автора текста, где учитываются индивидуально-личностные характеристики каждого автора.

С целью сбора базы данных для исследований была разработана утилита, в автоматическом режиме собирающая комментарии пользователей с интернет-форумов, работающих на основе систем управления сайтом phpBB и Invision Power Board. С её помощью на интернет-форуме http://forum.tomsk.ru для экспериментов было собрано 41 780 сообщений 30 авторов (15 мужчин и 15 женщин). Сообщения авторов выбирались таким образом, чтобы охватить как можно больше тем обсуждений (знакомства, политика, автомобили и т. д.). Все представленные авторы имеют большой «стаж» общения (более двух лет), что подтверждается весомой серией сообщений у каждого из них. Текст сообщений был очищен от всех вспомогательных метаданных, так или иначе способных повлиять на результаты исследования. Однако в дальнейших исследованиях авторами планируется использовать часть данной информации в качестве дополнительных атрибутов при идентификации (в частности BB-коды). Информация о корпусе представлена в табл. 1 и на рис. 1. Стоит отметить, что все собранные сообщения были обращениями к некоторому собеседнику.

Таблица 1. Средняя длина сообщения в символах

Отправитель и получатель	Средняя длина сообщения, символов	Количество сообщений
Мужчин мужчинам	57,3	10 406
Мужчин женщинам	61,5	8 179
Женщин мужчинам	62,9	8 351
Женщин женщинам	65,3	14 844
Мужчины	59,4	18 585
Женщины	64,1	23 195

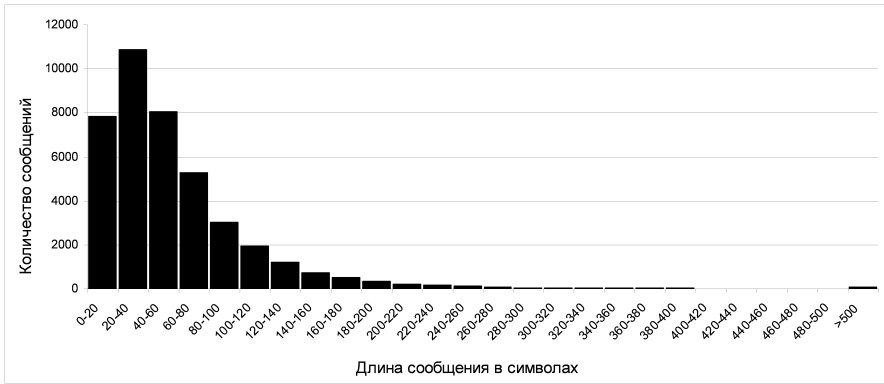

Рис. 1. Распределение длины сообщений в исследуемом корпусе

Из таблицы 1 видно, что средняя длина сообщения мужчин к мужчинам короче, чем сообщения мужчин к женщинам. Женщины пишут более длинные сообщения, чем мужчины независимо от пола получателя, с которым они общаются. Однако если получателем является также женщина, то такие сообщения, как правило, имеют наибольшую длину.

Как видно из графика на рисунке 1, большая часть собранных сообщений имеет длину не более 200 символов, поэтому тексты, состоящие из большего количества символов, не анализировались. Нижняя граница длины анализируемых сообщений была ограничена 20 символами. Этим условиям удовлетворяют 32 555 сообщений, т. е. около 78 % первоначального корпуса.

Для исследований использовалась программная система «Авторовед» [6], разработанная с целью статистического анализа текста на различных уровнях его организации и исследования характеристик текста задачах атрибуции текстов. Программная система успешно применялась в исследовательских целях для решения задачи идентификации автора литературных текстов и коротких сообщений [5, 7], а также для решения ряда частных практических задач. В частности для коротких текстов длиной до 100 символов удалось достичь точности 0,7 путем анализа частот наиболее частых слов русского языка, наиболее частых триграмм русского языка, униграмм символов, знаков препинания (одиночных и составных) и эмотиконов.

В качестве классификатора в настоящем исследовании используется машина опорных векторов (Support Vector Machine, SVM), математический аппарат которой был предложен В. Н. Вапником [8]. SVM может работать напрямую с векторным пространством высокой размерности без необходимости предварительного анализа и снижения количества измерений. Метод SVM изначально предназначен для классификации по двум возможным альтернативам, поэтому, как нельзя лучше, подходит для решения задачи определения пола автора. Для обучения моделей SVM применяется метод последовательной оптимизации (Sequential Minimal Optimization) [9], ядро выбрано линейное, параметр регуляризации C = 1, допустимый уровень ошибки — 0,00001.

Для выявления признаков текста, позволяющих определить пол автора, выполнялась следующая последовательность действий:

1. Из корпуса случайным образом извлекались тексты для обучения в количестве 5000 и тестирования в количестве 2500. Первая группа используется для обучения модели классификатора. Вторая — для проверки точности с помощью обученной модели.
2. Формирование вектора признаков для каждого из сообщений.
3. Приведение значений признаков в единый диапазон с помощью операции нормирования. Использовалось минимаксное нормирование в диапазон [-1..1].
4. Корректировка параметров классификатора, позволяющих обеспечить высокую разделяющую способность, путем обучения классификатора на нормированных векторах признаков группы обучающих текстов и проверки точности обученного классификатора на векторах признаков тестовой группы текстов.
5. Изменение перечня групп характеристик и/или признаков, составляющих группу, в случае, если изменением параметров классификатора достичь приемлемых результатов не удается.

Всего было исследовано порядка 50 различных признаков текста на уровне символов, слов и предложений, а также их сочетаний. Для каждой из характеристик было проведено по 20 описанных выше опытов. В качестве результирующей точности по данному признаку подсчитывалась средняя частота правильных классификаций. Результаты исследований представлены в табл. 2 (представлены характеристики, показавшие наилучший результат).

Таблица 2. Результаты экспериментов

Характеристика текста	**Точность**
униграммы символов	0,57
биграммы символов	0,51
триграммы символов	0,59
условные биграммы символов	0,53

Характеристика текста	Точность
служебные слова	0,58
распределение длин слов	0,57
знаки пунктуации и эмотиконы	0,68
словарный запас	0,52
ансамбль SVM	0,74

В результате исследований были определены характеристики короткого сообщения, имеющие преимущественное значение для использования в методиках определения пола автора. К ним можно отнести употребление человеком определенных сочетаний букв, служебных слов русского языка, знаков пунктуации, придание эмоциональной окраски высказыванию с помощью эмотиконов. Также в вопросе определения пола автора можно ориентироваться на длину слов в сообщениях.

Точность, превышающая 0,5, позволяет сделать вывод о принципиальной возможности определения пола автора короткого электронного сообщения на русском языке. Обучим модели SVM отдельно на каждой из групп признаков и объединим результаты классификации следующим образом — итоговым решением считается автор, выбранный большинством классификаторов. Использование такого ансамбля классификаторов позволило увеличить точность определения пола автора до 0,74. Это позволяет сделать вывод о целесообразности использования ансамблей классификаторов для принятия итогового решения и дальнейшего развития данной группы методов классификации в контексте решаемой задачи.

References

1. *Cheng N., Chen X.* et al. 2009. Gender Identification from E-mails. Proceedings of IEEE Symposium on Computational Intelligence and Data Mining : 154–158.
2. *Doyle J., Keselj V.* 2005. Automatic Categorization of Author Gender via N-Gram AnalySis. Proceedings of The 6th Symposium on Natural Language Processing, SNLP'2005, available at: http://web.cs.dal.ca/~vlado/papers/SNLP05J.pdf.
3. *Köse C., Özyurt Ö., Amanmyradov G.* 2007. Mining Chat Conversations for Sex Identification. Emerging Technologies in Knowledge Discovery and Data Mining Lecture Notes in Computer Science, 4819 : 45–55.
4. *Langer J., Jones V., McNabb M.* Gender Differences in Text Message Content, available at: http://www.jennalanger.com/academic/Langer-Jenna-Gender_dif_SMS_Content.pdf.
5. *Platt J. C.* 1999. Fast Training Support Vector Machines using Sequential Minimal Optimization : 185–208.
6. *Romanov A. S., Meshcheriakov R. V.* 2009. Text Author Identification by Support Vectors Device [Identifikatsiia Avtora Teksta s pomoshch'iu Apparata Opornykh Vektorov]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy

Mezhdunarodnoi Konferentsii “Dialog 2009” (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference “Dialog 2009”), 8 (15) : 432–437.

7. *Romanov A. S.* 2009. Programming System for Written Text Author Identification “Avtoroved” [Programmnaia Sistema dla Identifikatsii Avtora Pis’mennoi Rechi “Avtoroved”]. Khroniki Ob”edinennogo Fonda Elektronnykh Resursov “Nauka I Obrazovanie”, 7 : 7.
8. *Romanov A. S., Meshcheriakov R. V.* 2010. Short Message Author Identification with the Methods of Machine Learning [Identifikatsiia Avtorstva Korotkikh Tekstov Metodami Mashinnogo Obucheniia]. Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2010” (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference “Dialog 2010”), 9 (16) : 407–413.
9. *Vapnik V.* 1998. Statistical Learning Theory.
10. *Platt J. C.* 1999. Fast Training Support Vector Machines using Sequential Minimal Optimization : 185–208.