Опознание неизвестных пользователей Интернет-ресурсов по их техническим и лингвистическим аспектам

Аннотация. В последнее время наблюдается существенное возрастание числа пользователей различных Интернет-порталов. Некоторые из них не желают указывать достоверные сведения при регистрации на тех или иных ресурсах.

Фактор анонимности главным образом влияет на статистику противоправных действий в Интернете. К примеру, из года в год увеличивается количество анонимных и экстримисткских угроз. Ситуация усугубляется и тем, что на данный момент не существует действенных подходов и четких алгоритмов распознавания неизвестных юзеров. В связи с этим, идентификация анонимных пользователей Интернет-порталов становится важной и актуальной научной задачей.

В данной статье под идентификацией анонимных пользователей подразумевается процесс распознавания путем сопоставления набора их характеристик с характеристиками, которые были собраны ранее и которые уже имеются в соответствующей базе данных [1-5]. К таким показателям относят:

- лингвистические – стиль письменной речи автора текста;
- технические – тип операционной системы, IP-адрес;
- комбинированные характеристики.

В работе рассматривается подробно каждый вид, а также анализируется возможность использования разных методов классификации (способ логической регрессии, опорных векторов и др.) для решения поставленной задачи. В результате проведенных экспериментов стало очевидно, что применение технических и лингвистических характеристик в комплексе позволяет достичь определенной точности при идентификации анонимных пользователей Интернета.

Ключевые слова: авторство пользователей, распознавание неизвестных юзеров, идентификация анонимных пользователей, информационная безопасность, характеристика сообщений

Введение

Вопрос анонимности интернет-пользователей носит дискуссионный характер. Нет однозначного мнения относительно того, является ли возможность анонимной регистрации положительной или отрицательной чертой.

Необходимость в определении личности пользователя Интернет-порталов возникает в тех случаях, когда через Интернет были совершены действия, противоречащие закону. К таким противоправным действиям можно отнести сознательную анонимную дезинформацию других юзеров, либо комплекс действий, осуществляемый через переписку в Интернете и проводимый с целью организации преступления [6].

Способы решения задачи идентификации интернет-пользователей могут быть задействованы в следующих сферах: борьба с терроризмом (поиск опасных террористов при помощи их переписок в Интернете), компьютерная криминалистика (установление личности автора угрожающих сообщений) и прочее [1-13]. Актуальности данной темы обосновывается также большим количеством научных работ, посвященных решению задачи идентификации анонимных пользователей Интернета.

В рамках статьи дается следующее определение понятию «пользователь Интернет-портала» – это конкретное лицо, которое собственными действиями на сайте позволяет определить его характерные признаки при помощи анализа особенностей его письменного стиля общения и через установление типа технического средства, используемого им для доступа в Интернет. Цель работы заключается в оценивании возможностей применения

определенных наборов характерных признаков и способов классифицирования для идентификации анонимных интернет-пользователей.

Задача определения набора характеристик пользователей Интернета

Идентификация анонимных пользователей Интернет-порталов зачастую сводится к применению многоклассовой классификации. Принцип данной методики заключается в том, что существует:

- множество объектов (наборов характеристик юзеров): $F = \{f1, \dots fn\}$;
- множество самих юзеров: $U = \{u1, \dots uk\}$, $u_j = fk$.

Для определенного числа признаков $F' = \{f1, \dots, fm\} \subseteq F$ установлено их принадлежность конкретному объекту, то есть имеется некоторое множество выявленных связей «пользователь – характеристики». При этом главная задача заключается в том, чтобы определить кому из множества U относятся другие признаки: $F' = \{f_{m+1} \dots fn\} \subseteq F$.

Таким образом, для идентификации юзера следует сформировать следующий алгоритм $a: F \to U$, который позволял бы распределять по группам любой набор характеристик непосредственно из имеющегося множества – $fi \in F$. При этом U – включает в себя множество «характеристики – известные пользователи», F' – признаки, которые классифицируются, а Ftr – обучающая подборка.

Характерные особенности пользователей

Наиболее существенной задачей при идентификации юзеров выступает определение набора тех признаков (атрибутов), которые к нему относятся и которые позволят с максимальной точностью произвести классификацию. При этом количество возможных атрибутов достаточно велико. Наиболее простые из них это лингвистические и технические.

Лингвистические характеристики включают в себя, к примеру, показатель частоты использования слов определенной длины, а также другие более сложные признаки, которые требуют проведения семантического или синтетического анализа [2]. К техническим относят версию ОС, IP-адрес [4].

Не менее важной задачей считается разработка действенных методов, необходимых для идентификации пользователей, либо проведение классификации набора признаков, принадлежащих данному юзеру, при помощи сопоставления его характеристик с характеристиками известных пользователей. В контексте данной работы приведем ряд наиболее эффективных моделей для идентификации, позволяющих максимально точно классифицировать характеристики [1-13].

Таким образом, для достижения поставленной цели сгруппируем атрибуты юзера:

$$U = \{Ft, Fl\}, \quad (1)$$

где Ft – технические характеристики;
Fl – лингвистические.

В таблице 1 отражены используемые характеристики и указаны особенности их применения при идентификации пользователя.

При идентификации пользователя следует также учитывать тот факт, что для выхода в Интернет он может использовать несколько устройств, с юзером (u_i) также необходимо сопоставить некоторый стиль его текста на родном языке (fl) и прочие важные характеристики технических средств (ft_i). Таким образом, юзера можно отобразить следующим образом:

$$u_i = f_i = \{ft_i, fl_i\}, \quad u_i \in U, \quad f_i \in F, \quad ft_i \in Ft, \quad fl_i \in F_l \tag{2}$$

где f_i – общие характеристики пользователя;
F – множество возможных групп признаков юзера;
ft_i – критерий технических средств пользователя;
Ft – множество возможных значений атрибутов технических средств;
fl_i – лингвистический критерий текста;
F_l – множество возможных атрибутов текста.

Особенность технических характеристик заключается в том, что они не являются индивидуальными для пользователей, поскольку характеризуют только используемое им устройство. Анализ подобных признаков отображен в следующих работах [4, 11].

Технические характеристики также имеют следующий недостаток – в случае проведения расследования после осуществления преступного деяния, выявить технические атрибуты пользователя невозможно, так как сбор подобных сведений зачастую не ведется на сервере. В свою очередь лингвистические характеристики могут анализироваться в любое время – пока текст сообщений не удален из базы данных Интернет-портала.

Таблица 1 – Характеристики интернет-пользователя и специфика их применения при идентификации

Характеристики пользователя	**Особенности использования при идентификации**
Технические характеристики (F_t)	
Версия операционной системы Часовой пояс Версия пользователя Интернет Язык браузера по умолчанию Разрешение экрана	Показатели могут меняться и их весьма просто заменить с помощью заголовков HTTP-запроса, поэтому для достоверности лучше всего использовать активные элементы страниц.
IP-адрес	Показатели различимы на веб-сервере. Одновременно может быть использован несколькими физическими лицами в 4-ой версии. Активные элементы не используются в связи с незащищенностью ОС конечного пользователя.
Лингвистические характеристики (F_l)	
Общее количество символов (S) Частота буквенных символов Частота заглавных букв Частота цифр Частота пробелов Частоты управляющих символов Общее количество слов (W) Частоты размерности слов Частота коротких слов Средний размер слова Средний размер предложений в символах Средний размер предложений в словах	**Критерии текста как классификационный признак –** слова, имеющие высокие частотные характеристики, т.е. чаще всего используются автором. **Минимальный объем и содержательность текстов обучающей выборки.** Доступным является лишь ограниченный набор букв или слов для обработки. **Нарушенная гармония классов.** Неравномерность распределения текстовой

Частота коротких предложений Частота средних предложений Частота длинных предложений Частоты знаков препинания Частота применения ссылок Частота применения изображений Прочее	размерности обучающей выборки юзеров, способствующая получению неверного итога.

Описание эксперимента

Для проведения эксперимента по авторской идентификации был составлен блок текстов разной тематики на русском языке. При этом автор доподлинно был известен, а тексты дополнительно никак не обрабатывались. Такие исходные факторы позволяют создать максимально приближенные к жизни условия. Для обучающей подборки были отобраны последние 25 сообщений произвольного размера:

1) до 2700 символов – 73%;

2) 2700-5500 символов – 10%;

3) 5500-8200 символов – 6,5%.

Технические характеристики вместе с лингвистическими были отобраны таким образом, чтобы методы их подбора не противоречили принципам этики. Это создает некоторые ограничения при проведении эксперимента. Что касается получения MAC-адреса, то это не только неэтично, но и связано с некоторыми ограничениями применения активных элементов на странице ресурса.

Еще одним важным фактором является то, что пользоваться одним устройством может несколько человек. Таким образом, для классификации были отобраны следующие доступные технические характеристики:

- тип ОС;
- версия и язык браузера;
- разрешение экрана;
- часовой пояс;
- IP-адрес (версия 4).

В ходе эксперимента протестировали несколько алгоритмов α: $F \rightarrow U$, способных группировать характеристики из исходного множества признаков, а именно:

- метод логистической регрессии (LR);
- наивный байесовский классификатор (NB);
- многослойный перцептрон (NN);
- метод опорных векторов (Support Vector Machine, SVM).

Для анализа были использованы классификационные признаки, отраженные в таблице 1. Таким образом, чтобы оценить точность эксперимента, необходимо применить данный алгоритм на тестовой подборке характеристик и сопоставить полученный результат с ранее известным ответом.

Достигнутый эффект

На качество классификации непосредственное влияние имеют следующие факторы:

- тип применяемых признаков;
- количество юзеров;
- метод распределения характеристик.

Алгоритм тестировался на выборке, в которой было проставлено соответствие между пользователями и признаками. Численное оценивание алгоритма проводилось посредством оценки его точности (accuracy).

Точность анализа (A) определяется путем соотношения числа верно определенных наборов характеристик $(fi \epsilon F') - C_c$ к общему количеству классифицируемых наборов признаков C_t (величина подборки) [14]. Формула имеет следующий вид:

$$A = \frac{C_c}{C_t} 100\% \quad (3)$$

Для отдельного набора характеристик была проведена оценка качества выбранного ранее алгоритма, для этого полученные результаты сравнивали с результатами других алгоритмов. Результаты для лингвистических, технических и комбинированных характеристик отражены в таблице 2. Следует отметить, что результаты можно сопоставлять между собой, так как для каждой группы применялся один общий комплекс текстов.

Применение технических атрибутов считается традиционным подходом. В свою очередь, применение лингвистических характеристик в разрезе русского языка – это относительно новая задача, для решения которой требуется более углубленное изучение. Следует заметить, что это направление считается наиболее перспективным.

Из таблицы 2 и приведенного рисунка 1 можно отметить, что применение для анализа комбинированной совокупности характеристик позволяет значительно увеличить качество классификации. При этом ее точность составляет 90,4%.

Таблица 2 – Точность идентификации при использовании разных совокупностей характеристик автора

	Технические характеристики (Ft)	**Лингвистические характеристики (Fl)**	**Комбинированные характеристики (Ft + Fl)**
SVM	72,0	60,0	18,4
NN	72,1	61,6	90,4
NB	66,0	68,8	77,6
LR	68,9	79,2	89,6

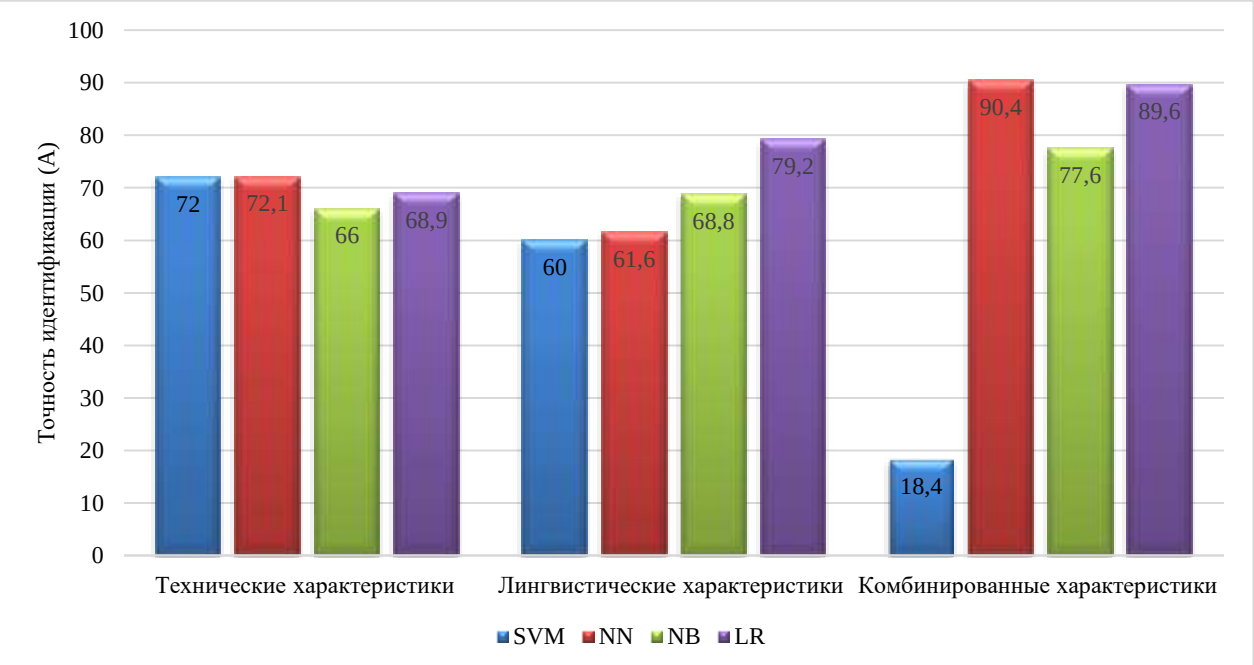

Рисунок 1 – Качество проведения идентификации юзеров при использовании разных совокупностей характеристик

Заключение

В данной статье для решения задачи идентификации интернет-пользователя был предложен действенный способ классификации, принцип которого заключается в анализе комбинированной совокупности характеристик автора. Для эксперимента были использованы следующие наборы характеристик: лингвистические (отражают характер письменной речи) и технические (характеризуют свойства технических средств, используемых пользователем для доступа в Интернет). Особенность данного подхода заключается в том, что его следует применять для классификации в условиях небольшого размера выборки.

Комплексное применение перечисленных признаков пользователей и предложенного алгоритма позволяет достичь желаемого результата при идентификации авторов текстов. Следует заметить, что метод опорных векторов имеет невысокую эффективность при использовании его на комбинированных характеристиках. В свою очередь, как показывают научные эксперименты, значительной эффективности идентификации можно достичь при использовании комбинированного комплекса характеристик в условиях ограниченной тестовой подборки.

Таким образом, высокая точность идентификации напрямую зависит от качества и информативности выбранных для анализа характеристик автора. Однако не решенным остается вопрос о том, какой из приведенных методов позволяет достичь наиболее точных результатов. Для установления истины требуется более детальное исследование.

Список используемой литературы

1. de Vel A., Anderson O., Corney M., Mohay G. Mining e-mail content for author identification forensics // SIGMOD Record. 2001. V. 30. N 4. P. 55-64.
2. Zheng R., Li J., Chen H., Huang Z. A Framework for Authorship Identification of Online Messages: Writing Style Features and Classification Techniques // Journal of the American Society of Information Science and Technology. 2006. V. 57. N 3. P. 378-393.
3. Iváncsy R., Juhász S. Analysis of Web User Identification Methods // International Journal of Computer Science. 2007. V. 2. N 3. P. 172-177.
4. Бессонова Е.Е., Зикратов И.А., Росков В.Ю. Анализ способов идентификации пользователя в сети Интернет // Научно-технический вестник информационных технологий, механики и оптики. 2012. № 6 (82). С. 128-129.
5. Романов А.С., Шелупанов А.А., Бондарчук С.С. Обобщенная методика идентификации автора неизвестного текста // Доклады Томского государственного университета систем управления и радиоэлектроники. 2010. № 1 (21). Ч. 1. С. 108-112.
6. Гвоздев А.В., Лебедев И.С., Зикратов И.А. Вероятностная модель оценки информационного воздействия // Научно-технический вестник информационных технологий, механики и оптики. 2012. № 2 (78). С. 99-103.
7. Abbasi A., Chen H. Applying Authorship Analysis to Extremist-group Web Forum Messages // IEEE Intelligent Systems. 2005. V. 20. N 5. P. 67-75.
8. Park T., Li J., Zhao H., Chau M. Analyzing writing styles of bloggers with different opinions // Proc. of the 19th Annual Workshop on Information Technologies and Systems (WITS 2009). Phoenix, Arizona, USA, 2009. P. 151-156.
9. Layton R., Watters P., Dazeley R. Authorship attribution for twitter in 140 characters or less // Second Cybercrime and Trustworthy Computing Workshop (CTC-2010). Ballart, VIC, Australia, 2010. P. 1-8.
10. Zheng R., Li J., Chen H., Huang Z. Authorship analysis in cybercrime investigation // Proc. of the 1st NSF/NIJ conference on Intelligence and security informatics (ISI'03). Berlin-Heidelberg: Springer-Verlag, 2003. P. 59-73.

11. Eckersley P. How Unique is Your Web Browser? // Lecture Notes in Computer Science. 2010. V. 6205. P. 10-18

12. Stamatatos E. A Survey of Modern Authorship Attribution Methods // Journal of the American Society for Information Science and Technology. 2009. V. 60. N 3. P. 538-556.

13. Nawrot M. Automatic Author Attribution for Short Text Documents // Lecture Notes in Computer Science. 2011. V. 6562. P. 468-477.

14. Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск. М.: Вильямс, 2011. 528 с.