УДК 004.056

Разработка модели пользователя сети Интернет для решения задачи лингвистической идентификации

М.С. Базанов, Университет ИТМО, Санкт-Петербург

Научный руководитель – к.т.н., доцент А. А. Воробьева, Университет ИТМО, Санкт-Петербург

Введение.

Современное общество невозможно представить без существования Интернета. Деловая переписка, научные исследования, общение по интересам и даже функционирование организаций – все это и многое другое невозможно полноценно представить без использования мировой сети.

Текстовые общение и информация в сети обезличены, в отличии от аудио и видео информации. По прочитанному тексту нельзя сразу достоверно установить его авторство. Преступник может обладать техникой жертвы, знать его пароли, владеть электронными идентификаторами. И тогда становится крайне затруднительно в короткие сроки установить факт подмены личности. От чужого имени он может совершать противоправные действия и оставаться анонимным.

Одним из наиболее перспективных и устойчивых к фальсификации методов идентификации является биометрическая идентификация. Одной из её разновидностей является лингвистическая идентификация. Данный вид идентификации может применяться в системах разграничения доступа, в компьютерной криминалистике, в системах электронной почты и обмена сообщениями, в системах по определению уникальности текстов (например, «Антиплагиат») как дополнительная функция, а также во многих других сферах и технологиях, где необходимо идентифицировать пользователей.

Цель.

Разработать модель пользователя сети Интернет для решения задачи лингвистической идентификации.

Базовые положения исследования.

Для решения задачи лингвистической идентификации пользователей необходимо разработать модель пользователя, которая будет применяться в процессе самой идентификации. Данная модель будет являться своего рода «цифровым профилем» пользователя, которая будет характеризовать его по различным выбранным параметрам.

Так как каждый человек обладает уникальным набором лингвистических характеристик, то параметры модели для каждого пользователя будет уникальна, что позволит использовать её для идентификации.

При составлении модели необходимо учитывать множество факторов, которые существенно могут влиять на письменную речь пользователей. В их число может входит как внешние факторы (состояние окружения), но также и общее психологическое и физическое состояние, а также замыслы и желания.

Важными аспектами в составление модели является анализ социальных сетей пользователя и других коммуникационных платформ в сети Интернет и анализ его собственных сообщений, авторство которых, условно, известно достоверно. Анализируя общедоступную информацию, такую как «никнеймы», адреса электронных почт, должности, имена, указанные на сайтах и т. д., можно определить многие черты его характера, особенностей письменной речи, что позволит, анализируя его сообщения, идентифицировать его по каким-либо ключевым словам, которые характерны для него в силу профессии, характера, возраста и т.д. Эта информация может быть также использована в поиске

остальных профилей человека на различных сайтах в сети Интернет и дополнительной информации, которая поможет в лингвистической идентификации пользователей.

Безусловно, такая идентификация связана с большим числом проблем и трудностей, которые возникают во время анализа текстов, сбора необходимой информации. Среди основных можно выделить сложность самого процесса идентификации и её не полную точность. Однако будущие исследования в этой области должны помочь решить эти и другие проблемы, чтобы обеспечить высокий уровень надежности лингвистической идентификации.

Промежуточные результаты.

В работе описываются процессы сбора необходимой для построения модели информации, определения необходимых и возможных параметров модели и построения непосредственно самой модели. Проведен небольшой сравнительный анализ имеющихся моделей для русского языка.

Основной результат.

В ходе работы удалось построить модель пользователя сети Интернет для решения задачи лингвистической идентификации пользователей, которую в дальнейшем можно использовать для решения различных прикладных задач в различных сферах человеческой деятельности.