Технологии выявления и очистки персональных данных открытых источников*

В.А. ИВИЧЕВ, Т.В. ИГНАТОВА, ООО«Медиалогия», Москва.
E-mail: vivichev@mlg.ru, tignatova@mlg.ru

В статье рассмотрены основные задачи выявления персональных данных в массивах неструктурированной информации, а также обезличивания персональных данных. Их решение позволит осуществить качественный скачок в вопросах оперативного выявления нарушений федеральных законов «О персональных данных», «О средствах массовой информации» (о СМИ), «Об информации, информационных технологиях и защите информации». Основные достоинства предлагаемых в статье технологий лежат в плоскости извлечения из неструктурированной информации средствами компьютерной лингвистики объектов (персона, бренд, организация, географическое понятие), их атрибутов и связанных с объектами фактов, а также дальнейшей лингвистической и аналитической обработки извлеченной информации. Такие технологии могут использоваться в информационно-аналитических системах, предназначенных для оперативного и точного выявления персональных данных, в интересах как органов государственной власти, так и коммерческих организаций и физических лиц.

Ключевые слова: персональные данные, обезличивание перснальных данных, прозрачность и доступность информации, выявление нарушений, защита информации, оценка сообщений открытых источников информации, выявление фактов, информационная система, федеральный закон «О персональных данных» (№ 152-ФЗ), федеральный закон «О средствах массовой информации» (о СМИ) (№ 2124-1-ФЗ), федеральный закон «Об информации, информационных технологиях и защите информации» (№ 149-ФЗ)

Проникновение сети Интернет в повседневную жизнь миллионов людей по всему миру породило вместе с невиданными ранее возможностями ряд новых проблем. Так, например, повседневная обработка персональных данных, определенных законом как любая информация, относящаяся прямо или косвенно к определенному или определяемому физическому лицу (субъекту персональных данных), порождает необходимость защиты данных, а также выявления случаев и источников нарушения соответствующих законов.

* Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации. УДК 004.031.42.

Федеральный закон «О персональных данных» регулирует, в частности, отношения между субъектом персональных данных и оператором[1]. Однако нарушения закона (несанкционированное распространение персональных данных) носят массовый характер ввиду того, что точное и оперативное выявление фактов публикации персональных данных конкретного лица в огромном потоке открытой информации – очень трудоемкая и технически сложная задача, а отечественная судебная практика пока недостаточна: в ряде случаев определение подсудности дела затруднено[2], и даже если вина установлена, то штрафные санкции не эквивалентны выгоде нарушителя. Часто виновных лиц (владельцев ресурсов, на которых незаконно размещены персональные данные) к ответственности не привлекают в связи с истечением установленных сроков давности (3 месяца)[3].

Проблемы соблюдения конфиденциальности персональных данных наиболее остро волнуют граждан и обсуждаются в СМИ. Так, согласно отчету федеральной службы по надзору в сфере связи, информационных технологий и массовых коммуникаций (Роскомнадзор), за 2011 г. из 3365 обращений граждан 3214 (82%) составили жалобы о нарушениях операторами их прав и законных интересов в области персональных данных[4]. Вслед за скандалами вокруг утечек персональных данных интернет-магазинов, МВД, ГИБДД и т.д. последовало нормативное закрепление процедур обезличивания персональных данных операторами[5].

Постановление Правительства РФ от 21.03.2012 г. № 211 (п. 1, абзац «з») предписывает следующую ответственность оператора: *«...согласно требованиям и методам, установленным уполномоченным органом по защите прав субъектов*

---

[1] Государственный орган, муниципальный орган, юридическое или физическое лицо, самостоятельно или совместно с другими лицами организующие и (или) осуществляющие обработку персональных данных, а также определяющие цели обработки персональных данных, состав персональных данных, подлежащих обработке, действия (операции), совершаемые с персональными данными.

[2] URL: http://www.pd.rsoc.ru/faq/faq4.htm (Дата обращения: 15.01.2013).

[3] Отчет о деятельности Уполномоченного органа по защите прав субъектов персональных данных за 2011 год. URL: http://www.rsoc.ru/docs/Otchet_Upolnomochennogo_organa_za_2011_g.doc (Дата обращения: 15.01.2013).

[4] Там же.

[5] Постановление Правительства Российской Федерации от 21.03.2012 г. № 211.

*персональных данных, операторы (*прим. авт.*) осуществляют обезличивание персональных данных, обрабатываемых в информационных системах персональных данных, в том числе созданных и функционирующих в рамках реализации федеральных целевых программ».*

Потребность в эффективных и высокотехнологичных средствах для автоматизированных выявления и очистки (обезличивания) информации от персональных данных в огромном непрерывном потоке информации открытых источников испытывают практически все компании и организации, которые обрабатывают или готовят к публикации информацию, содержащую данные, подлежащие защите в соответствии с федеральным законом «О защите персональных данных» № 152-ФЗ. Но на российском рынке отсутствуют готовые продукты, полностью охватывающие весь спектр указанных задач. Для их решения предлагаются технологии, речь о которых пойдет в данной статье.

Используемые технологии

По мнению экспертов, эффективность использования лингвистических технологий в процессе обнаружения и обезличивания персональных данных может достигать 95%[6]. Поэтому для реализации информационной системы выявления и обезличивания персональных данных были разработаны и применены следующие, в том числе и лингвистические, технологии:

- мониторинга открытых источников информации (веб-сайты органов государственной власти, печатные и интернет-СМИ, блоги и форумы);
- лингвистической обработки неструктурированной информации открытых источников;
- выявления в тексте упоминаний персон и связанной персональной информации;
- автоматического устранения из текста персональных данных с сохранением смысловой целостности текста (обезличивание персональных данных).

[6] URL: http://ru.wikipedia.org/wiki/Information_Protection_and_Control# (Дата обращения: 15.01.2013).

Если методы мониторинга, сбора информации из открытых источников широко известны, то о применении лингвистических технологий на этапе обработки собираемой неструктурированной информации стоит сказать отдельно.

Лингвистическая обработка информационной системы выявления и обезличивания персональных данных включает следующее.

1. Выделение информационного объекта и ранжирование важности его упоминания в тексте сообщения (главная, второстепенная или эпизодическая роль). Извлекаемая информация также используется как вспомогательная для оценки потенциального ущерба от публикации персональных данных при расчете индекса качества[7].

2. Рубрицирование текстов, в ходе которого обрабатываемые тексты относятся к одной из трех категорий (допускается их расширение): открытые, частные, закрытые. Это позволяет упростить критериальный пользовательский поиск нарушений и упорядочение представления данных в интерфейсе пользователя информационной системы. Очевидно, что публикация сведений о должности, месте работы, образовании, за редким исключением, не несет никакой угрозы, как правило, такого рода данные открыты. Напротив, серия и номер паспорта, адрес проживания персоны, сведения о доходах и т.п. – закрытая информация. Пользователи информационной системы могут запрашивать интересующий их набор типов персональных данных либо выбирать объединяющий класс персональных данных и получать сводку об их публикации.

3. Жанровая классификация текстов требуется, например, для анализа статистики публикации персональных данных в разрезе жанра (интервью, аналитическая статья, запись в блоге и т.д.).

4. Выявление групп информационных событий и автоматическая кластеризация поступающих информационных материалов. Кластеризация применяется для выявления новых рубрик персональных данных. Например, информация о доходах чиновников, в отличие от информации о доходах остальных категорий населения, является открытой. Обработка

[7] Индекс качества – расчетный индекс, отражающий качественную оценку отношения к заданному объекту в тексте открытого источника.

опубликованных на сайтах ведомств деклараций о доходах чиновников сформирует особый кластер, что позволит исключить декларации из результатов поиска документов с закрытыми персональными данными. Таким образом, система способна учитывать множество сложных условий для классификации персональных данных.

5. Расчет индекса качества для выявленных информационных объектов. Индекс качества позволяет оценить степень потенциального ущерба от публикации персональных данных в том или ином источнике для субъекта персональных данных.

Остановимся подробнее на технологиях выявления и обезличивания персональных данных. Задача выявления персональных данных решается поэтапно путем лингвистической обработки материала – текста (статьи, сообщения, документа) открытого источника. Комплексный разбор текста лингвистическим процессором информационной системы выявления и обезличивания персональных данных решает задачу разбиения текста на лексемы[8] и построения семантической сети, содержащей все сущности, найденные в тексте (наименования предметов и лиц, действий и признаков, связанных различными типами синтактико-семантических связей) по результатам разбора, морфологического анализа текста и выявления в нем элементарных сущностей (предложений, фраз, слов, символов)[9].

Далее специальный программный модуль выделяет в «разобранном» тексте факты присутствия персональных данных (сущностей, выявленных на первом этапе обработки, связанных с персонами), сравнивает готовые шаблоны фактов персональных данных с семантической сетью разобранного текста и при определенном совпадении – регистрирует присутствие персональных данных в тексте.

При обнаружении факта упоминания персоны в тексте выделяются следующие признаки:

- дата и место рождения персоны (полная дата рождения или частичное указание – год, месяц или определенный период в прошлом);

---

[8] Термин, предложенный А.М. Пешковским. *Пешковский А.М.* Методика родного языка, лингвистика, стилистика, поэтика. – Л., 1925.

[9] *Ермаков А.Е.* Извлечение знаний из текста и их обработка: состояние и перспективы //Информационные технологии. – 2009. – № 7.

- адрес места жительства (любое указание на регион проживания, город или более точный адрес);
- семейный статус (указание на текущий статус, наличие или отсутствие семейных отношений в прошлом);
- наличие детей;
- профессия, место работы (любое указание на род деятельности и место работы в прошлом и настоящем);
- оценка имущественного положения (указание на наличие имущества любого вида);
- информация о доходах (любая информация, указывающая на наличие доходов, источники и размеры доходов);
- информация об образовании (место обучения, вид образования, наличие документов, подтверждающих образование, даты обучения персоны).

Технология обезличивания персональных данных позволяет кодировать выявленные в тексте персональные данные с сохранением смысловой целостности обработанного текста. Это необходимо, например, при публикации судебных решений, для защиты от разглашения конфиденциальной информации в случае утечек баз данных операторов персональных данных и т.д.

Для сохранения смысловой целостности текста в рамках технологии реализована автоматическая замена упоминания всех персон в тексте на соответствующие кодовые обозначения: Персона 1, Персона 2, ..., Персона N (далее – Персона i). Кодовые обозначения сохраняются при замене упоминаний персоны по всему тексту, присвоение ей нескольких кодовых значений исключено. Все выявленные сопутствующие персональные данные аналогичным образом заменяются в тексте на кодовые обозначения, а именно:

- дата и место рождения персоны заменяется на дату рождения Персоны i и/или место рождения Персоны i;
- адрес места жительства – на адрес Персоны i;
- семейный статус – на кодовое обозначение в случае, если в данном тексте семейный статус используется для установления отношений между двумя персонами, в противном случае упоминание семейного статуса исключается из текста. Если в тексте упомянуты две персоны и их семейный статус, то упоминание также кодируется (Персона i находится в родственных отношениях с Персоной j);
- наличие детей – на кодовое обозначение в случае, если в данном тексте наличие детей используется для установления отношений между двумя персонами, в противном случае

упоминание семейного статуса исключается из текста. Если в тексте упомянуты две персоны и их семейный статус, то кодируется аналогично предыдущему случаю (Персона i находится в родственных отношениях с Персоной j);

- профессия, место работы – любое указание на род деятельности персоны и места работы (в прошлом и настоящем) – на кодовое обозначение – Род деятельности Персоны i.

Данная технология позволяет получать тексты с сохранением их первоначальной смысловой целостности, но с обезличенными персональными данными в структурированном виде (реляционная база данных, файл форматов xml, xls и т.д.). Таким образом, обратная процедура декодирования, т.е. возвращения персональных данных, но уже в структурированный и лингвистически размеченный текст, с одной стороны, не является технически сложной, с другой – открывает **широкие возможности для дальнейшей аналитической обработки, поиска и публикации такой информации**.

Архитектура информационной системы

Архитектура информационной системы формируется на основе анализа предъявляемых к системам данного класса функциональных требований (определяются потребностями потенциальных пользователей) и требований к качеству выявления и обезличивания персональных данных. В состав системы включены следующие компоненты (модули).

Модуль сбора, предварительной обработки и загрузки материалов открытых источников предназначен для поиска и загрузки (по расписанию) текстового содержимого интернет-источников, указанных в соответствующем справочнике системы; выделения и сохранения ссылок из их текста и дальнейшего обхода и загрузки связанных страниц источника по этим ссылкам, а также загрузки с целью дальнейшего выявления и обезличивания персональных данных текстовых документов из пользовательского интерфейса системы.

Данный модуль позволяет осуществить сбор информации из открытых источников следующих типов: веб-сайты, официальные печатные и интернет-СМИ, открытые интернет-блоги и форумы.

Модуль сбора обеспечивает решение следующих задач:

- сбор информации из открытых источников;
- предварительную обработку загруженных текстов – извлечение данных, необходимых для определения границ слов, предложений и параграфов в тексте с учетом переносов и вариации написания слов с использованием дефиса и т.д.;
- выделение из потока текстов, соответствующих заданному в настройках системы языку (кодовой странице);
- фильтрация дублей по рассчитываемой контрольной сумме;
- передача текстов для дальнейшей обработки (лингвистический анализ, разбиение текста на лексемы, извлечение из материалов необходимых служебных атрибутов, выявление объектов и персональных данных и т.д.);
- запись в базу данных системы всех результатов и истории работы процедур обработки текстов.

Модуль выделения и ранжирования объектов, определения роли объекта в тексте предназначен для определения роли (главная, второстепенная, эпизодическая) оцениваемого объекта в тексте для получения вспомогательной информации, необходимой, например, для оценки потенциального ущерба от публикации персональных данных при расчете индекса качества.

Модуль обеспечивает выделение в тексте и классификацию объекта, установление одного из трех возможных вариантов роли выделенного в тексте объекта на основании позиции в тексте, его типа, частоты упоминания объекта в тексте, а также запись в базу данных системы значений соответствующих атрибутов.

Модуль рубрицирования предназначен для отнесения обработанных материалов к одной из категорий персональных данных (закрытые, частные, общедоступные). Он упрощает обнаружение наиболее потенциально опасных (с точки зрения публикации персональных данных и возможных репутационных рисков) материалов, определяет принадлежность текста к рубрике на основе типов персональных данных, выявленных в нем.

Данный модуль обеспечивает категоризацию – отнесение обнаруженных персональных данных (в зависимости от степени потенциальной угрозы безопасности персоны) к группе закрытых, частных, общедоступных, а также запись в базу данных системы значений соответствующих атрибутов.

К категории *закрытых* относятся тексты, содержащие:

- паспортные данные;
- данные идентификационных документов (водительское, пенсионное удостоверения, полисы ОМС и пенсионного страхования);
- номер и прочие атрибуты пластиковой карты (CVV, дата окончания срока действия);
- доходы, имущество (не для госслужащих);
- данные о состоянии здоровья (о хронических заболеваниях), интимной жизни;
- номер автомобиля;
- домашний телефон;
- адрес регистрации, проживания;
- адреса находящихся в собственности объектов недвижимости.

К категории *частных* относятся тексты, которые содержат условно закрытые данные, доступные на сайтах поиска работы, объявлений, знакомств, через телефонный справочник, но при этом не позволяют установить местонахождение объекта (в нерабочее время), членов его семьи, связаться с ними. А также данные, публикация которых не представляет угрозы для репутации либо обязательна в силу служебных обязанностей – дата и место рождения, реквизиты банковского счета, адрес личной почты и номер личного мобильного телефона, гражданство, ОГРНИП, ИНН, данные документов об образовании, данные о расовой, национальной принадлежности, политических взглядах, религиозных или философских убеждениях, сведения о семейном положении, наличии детей.

К категории *общедоступных* относятся тексты, не позволяющие установить местонахождение объекта (в нерабочее время), членов его семьи, связаться с ними, а также данные, публикация которых не представляет угрозы для репутации объекта либо обязательна в силу служебных обязанностей (за исключением данных, позволяющих однозначно идентифицировать персону, а также сведений о расовой принадлежности и религиозных убеждениях). Это такие данные, как ФИО, пол, должность, место работы, номер служебного телефона, адрес служебной почты, образование (за исключением данных документов об образовании), сведения о воинском учете (за исключением данных об учетных документах), профессия, состояние здоровья, если речь не идет о хронических заболеваниях, сведения о доходах и имуществе (для госслужащих).

Модуль жанровой классификации предназначен для отнесения текста к определенной группе по жанру, определяемому

с помощью лингвистических технологий: новость, интервью, аналитика, аналитическая статья, TV, ток-шоу, законодательство, пресс-релиз, публицистика, отзывы, рейтинги, прочее.

Модуль обеспечивает определение жанра текста на основании его содержимого и запись в базу данных системы значений соответствующих атрибутов.

Модуль расчета индекса качества позволяет осуществлять комплексную оценку качества освещения объекта в тексте. Реализованный расчет системы индексов позволяет оперативно оценивать степень потенциального ущерба для упоминаемой в тексте персоны от публикации ее персональных данных. Индекс качества рассчитывается с использованием следующих данных:

- влиятельность открытого источника (рассчитанная на основе оперативно обновляемых данных о его цитируемости);
- номер полосы (для текстов печатных СМИ);
- размер текста;
- наличие иллюстрации;
- роль объекта в статье;
- наличие цитат объекта в статье;
- характер упоминания объекта (негатив или позитив).

Модуль позволяет произвести расчет индекса качества для выявленного в тексте объекта и запись в базу данных системы значений соответствующих атрибутов.

Модуль выявления персональных данных предназначен для выделения в тексте персональных данных, включая обнаружение фактов упоминаний персон и связанной с ними дополнительной персональной информации. В результате семантической обработки данных выявляются также связи между персонами, если они есть.

Данный модуль позволяет определить дату и место рождения, адрес и семейный статус персоны, наличие детей, а также профессию, место работы, информацию о доходах и об образовании и произвести запись в базу данных системы значений соответствующих атрибутов.

Модуль очистки персональных данных в тексте сообщения предназначен для автоматической замены в тексте выявленных персональных данных на кодовые обозначения с сохранением смысловой целостности исходного текста.

Модуль обеспечивает:

- подбор кодового слова или словосочетания из соответствующего справочника замен в соответствии с замещаемым элементом персональных данных (например, фраза *«Беляев А.Ф., проживающий по адресу: ул. Солянка, д. 1/2, кв. 47»* будет заменена на фразу *«[Персона], проживающая по адресу: [Адрес]»*);
- присвоение фразе замены порядкового номера. Так, если в примере из предыдущего пункта *Беляев А.Ф.* является второй упоминаемой в тексте персоной, а адрес встречается впервые, замена примет вид: «*Беляев А.Ф., проживающий по адресу: ул. Солянка, д. 1/2, кв. 47 – [Персона_2], проживающая по адресу: [Адрес_1]*»;
- запись в базу данных системы значений соответствующих атрибутов с сохранением связей – *исходное значение-замена*.

Модуль поиска и фильтрации текстов (исходных и обработанных статей, публикаций записей блогов) позволяет производить в базе данных системы:

- контекстный поиск по массиву исходных текстов, в том числе по указанной в качестве критерия части слова;
- расширенный поиск с возможностью выбора одного или нескольких параметров;
- фильтрацию результатов поиска по одному или нескольким параметрам.

Модуль получения отчетов и экспорта информационных сообщений предназначен для формирования аналитической отчетности на базе лингвистически обработанной информации с возможностью построения рейтингов по количеству выявленных нарушений в части персональных данных, фильтрации и выбора различной глубины ретроспективы и различных аналитических разрезов. Он также позволяет осуществлять экспорт результатов работы системы в файл.

Модуль обеспечивает:

- генерацию конечных отчетов одной из категорий: рейтинги, динамика показателей, региональные показатели, кластеры на основании информации, хранящейся в базе данных системы для их дальнейшей публикации в соответствующих разделах системы или экспорта результатов;
- экспорт отчетов, представленных в виде табличных данных, в файл формата xls;
- экспорт исходных и обработанных сообщений, содержащих персональные данные, в файл одного из форматов: doc, docx, rtf, xls, xlsx, txt, xml, pdf.

Модуль администрирования системы предоставляет пользователю, наделенному соответствующими полномочиями, инструментарий гибкой настройки и аудита системы, управления справочниками, словарями, журналами, учетными записями и правами пользователей, расписаниями работы служб и т.д. Он решает следующие задачи:

- контроль на основе расписания, заданного для каждого отрытого источника, своевременности поступления материалов в систему;
- управление расписанием поставки материалов, оповещениями о поступлениях;
- настройка *модуля сбора, предварительной обработки и загрузки материалов открытых источников*, включая ведение списка источников, установку порога глубины обхода интернет-источников и типов загружаемых и исключаемых файлов, адресов интернет-страниц для проверки обновлений содержимого открытого источника, определение дисковой квоты для источников и т.п.
- управление справочниками и словарями системы;
- просмотр журнала аудита системы с возможностью фильтрации сообщений по нескольким атрибутам.

***

Разработанная на основе описанных в статье технологий информационная система выявления и обезличивания персональных данных прошла экспериментальные испытания. **Результаты испытаний показали высокую точность выявления и обезличивания персональных данных в непрерывном потоке неструктурированной информации открытых источников в режиме реального времени на следующих объемах: 4582 открытых источника, в том числе наиболее популярные интернет-блоги и форумы, печатные и интернет-СМИ, сайты органов государственной власти и т.д. Только в процессе тестирования и испытаний системы в них автоматически выявлено более 122 тыс. текстов, содержащих публикации персональных данных.**