УДК 004.91

М. Д. Кузнецов, В. С. Мядзель
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина)

Е. С. Новикова
Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

Применение методов интеллектуального анализа текста для исследования согласий на использование персональных данных

Соглашения об использовании персональных данных, размещенные на веб-сайтах компаний, предоставляют пользователям информацию о том, какие персональные данные собираются, как они обрабатываются и каким третьим лицам они передаются. Однако в большинстве случаев соглашения написаны сложно, их содержание нечетко и непрозрачно. В статье авторы исследуют возможности двух различных подходов к анализу текста для исследования политик безопасности. Для выявления различных сценариев использования персональных данных предлагается использовать латентно-семантический, а для установления связей между элементами сценария – морфологический анализ. Морфологический анализ текста позволяет построить логические цепочки, характеризующие сценарии обращения с персональными данными. Авторы применили выбранные алгоритмы на наборе размеченных документов, собранных в рамках проекта Usable Privacy Project. Полученные результаты показали, что рассмотренные подходы могут применяться для решения поставленной задачи. Например, было показано, что для каждого сценария использования персональных данных можно построить некоторый фиксированный набор семантических моделей, с помощью которых можно оценивать наличие этого сценария в тексте политики безопасности.

Анализ текста, латентно-семантический анализ, поиск синонимов, контекстно-свободные грамматики, обработка естественного языка, соглашения об использовании персональных данных, морфологический анализ

В настоящее время сбор и обработка персональных данных пользователей широко применяется при предоставлении цифровых услуг, их персонализации и улучшении. Персональными данными считаются любые данные, «относящиеся к идентифицированному или идентифицируемому физическому лицу» [1]. Таким образом, персональные данные – это не только биометрическая информация, данные о состоянии здоровья человека, но и фотография абонента услуги, его местонахождение, информация о приложении и устройстве, с помощью которых можно было бы его отследить. Несколько массовых утечек персональных данных за последнее десятилетие привели к ужесточению требований законодательства во многих странах мира. В настоящее время требуется, чтобы все персональные данные обрабатывались безопасным и прозрачным для субъекта образом при его явно выраженном согласии.

Политики конфиденциальности поставщиков услуг и онлайн-соглашения пользователей вебсервисов – единственный законный способ сообщить пользователям информацию о том, какие персональные данные собираются, как они обрабатываются и каким третьим лицам они передаются. В большинстве случаев эти документы написаны достаточно сложным языком, и пользователи чаще всего не читают их. В настоящее время ситуация такова, что юридические требования соблюдаются производителями продукции и поставщиками услуг, но пользователи дают свое согласие без четкого понимания того, что происходит с их персональными данными. Это приводит к тому, что пользователи не знают о рисках, возникающих в результате обработки их конфиденциальной информации при использовании определенной услуги или устройства. Доста-

точно ярким примером может служить ситуация с использованием электронного пропуска в Москве во время эпидемии COVID-19 весной 2020 г. [2]. Чтобы получить электронный пропуск, позволяющий перемещаться по городу во время принудительной самоизоляции населения, пользователь должен был подписать согласие, которое включало пункт о передаче персональных данных третьим лицам в рекламных и маркетинговых целях на срок до десяти лет. Безусловно, данная ситуация может оцениваться пользователями услуги как некритичная, поскольку речь идет только о контактных данных пользователя, но необходимо понимать, что на месте контактных данных могут находиться финансовые данные, данные о здоровье, биометрические данные, и последствия их использования могут быть непредсказуемыми и даже трагичными для их владельцев.

Таким образом, важно разработать методы, повышающие прозрачность политик конфиденциальности, включая онлайн-соглашения пользователя на использование персональных данных, размещенных в Интернете, и позволяющие оценить риски конфиденциальных данных, связанных с использованием данной услуги или устройства. Авторы считают, что ключевым моментом в улучшении понимания потребителями онлайн-соглашения пользователей служит предоставление данной информации в четком и структурированном формате. В данной статье в качестве возможных решений для автоматического анализа политик безопасности рассматриваются несколько подходов – латентно-семантический анализ, позволяющий сформировать набор тем, представленных в документе, и морфологический анализ текста, позволяющий установить лексическую связь между словами для построения детального описания сценариев использования персональных данных. Инструменты анализа текста позволяют улавливать важные фрагменты текста и особенности онлайн-соглашения пользователя. Есть несколько успешных решений этой проблемы с использованием машинного обучения, например [3] и [4]. Но подходы, представленные в этой статье, проще и не требуют больших наборов данных, поскольку полагаются только на статистические параметры текста и правила грамматики.

Обычно согласие на использование персональных данных содержит описание нескольких сценариев их использования – сбор и обработку первыми лицами, передачу персональных данных третьим лицам, их удержание, обеспечение безопасности, описание способов получения к ним доступа и механизмы для управления ими, ориентированные на владельцев, описание механизмов уведомления в случае утечки персональных данных и изменения содержания политики конфиденциальности [5]. Возможность выделить несколько сценариев использования персональных данных привела к идее использования латентно-сематического анализа для построения семантических моделей тем, которые представлены в политике безопасности. В основе идеи использования морфологического анализа текста лежит следующее предположение: для каждого сценария использования данных можно определить, какие данные собираются, какова цель и какова правовая основа для такой обработки данных. В таком случае использование анализа частей речи кажется закономерным выбором, поскольку эта информация содержится в предложениях и на ее основе можно делать выводы о субъектах данных, объектах данных, а также о связях между ними. В статье представлены предварительные результаты тестирования этих двух подходов для решения проблемы обеспечения «прозрачности» соглашения пользователя на обработку его персональных данных.

Помимо данного раздела статья содержит еще три. В следующем разделе обсуждаются особенности латентно-семантического анализа для определения сценариев использования персональных данных. Третий раздел посвящен контекстно-свободному грамматическому подходу к анализу текста, который сценариев использования персональных данных. В последнем разделе обсуждаются полученные результаты и возможности предложенных методов по сравнению с моделями анализа текстов, основанными на обучении с учителем.

Применение латентно-семантического анализа к исследованию текста политики безопасности. Современные методы кластеризации текстов позволяют аналитику с высокой точностью определять тематику текстов. Однако большинство этих методов в качестве входных данных обычно принимают тексты, которые посвящены достаточно отличающимся темам. Очевидно, что текст можно анализировать дважды: в первый раз сгруппировать тексты по темам, а во второй раз выполнить более детальное разделение их по подтемам. Такой подход может быть использован для более точного определения абзацев до-

кумента с точки зрения их особенностей и аспектов сценария использования персональных данных. Однако следует отметить, что латентно-семантический поиск сильно зависит от глобального текстового контекста с потерей информации о локальных контекстных отношениях между словами.

Латентно-семантический анализ [6] основан на операции сингулярного разложения матрицы

$$A = U S V^{\mathrm{T}},$$

где A – это матрица частот появления слов в документах. Используя упомянутое разложение, вычисляются ортонормированные матрицы U, V и диагональная матрица S, значения которых сингулярны для A. После разбиения матрицы A на три компонента матрица U содержит n-мерные векторы, которые можно интерпретировать как элементы семантической модели документа. На основе этих моделей документы могут быть разбиты на кластеры или группы текстов, обладающих схожей тематикой.

Применение латентно-семантического анализа требует предварительной обработки входных данных. Обычно используемая процедура очистки данных включает удаление гиперссылок, знаков препинания и т. д. В предлагаемом подходе авторы решили разбивать тексты политик безопасности на параграфы, так как при их непосредственном прочтении было выявлено, что чаще всего каждый параграф текста содержит описание какого-либо одного сценария использования данных. Было также решено, что другие структурные элементы – таблицы, списки и т. п., будут обрабатываться в рамках параграфа, в котором они размещены. Каждый параграф преобразовывался в массив слов, из которого удалялись наиболее частые, но не очень значимые слова (стоп-слова). Кроме того, авторы также применили операцию выделения корней (стемминг), чтобы рассматривать только базовые части слов, образованных от них.

В экспериментах были протестированы две модели векторизованного представления текста – модель «мешок слов» (bag-of-words) и модель TF-IDF.

Модель «мешок слов» представляет документ в виде матрицы (рис. 1). Здесь характеристики слов подсчитываются и сопоставляются с параграфами, в которых они встречаются: d – параграф, w – слово, count – функция подсчета частоты употребления слова.

Слова

Параграфы	w_1	. . .	w_n
d_1	count (w_1, d_1)	. . .	count (w_n, d_1)
. . .	. . .	. . .	. . .
d_n	count (w_1, d_n)	. . .	count (w_n, d_n)

Частоты слов в параграфах

Рис. 1

Эксперименты с латентно-семантическим анализом проводились на наборе документов, который включает в себя 115 онлайн-согласий на использование персональных данных [3]. Документы размечены вручную, и каждый параграф отнесен к одному или нескольким способам использования персональных данных. Необходимо отметить, что авторами набора данных выделено 10 сценариев использования – сбор данных первыми лицами, передача данных третьим лицам, механизмы защиты персональных данных, сроки удержания персональных данных, механизмы доступа, редактирования и удаления персональных данных, механизмы контроля сбора данных, уведомление в случае изменения политики безопасности, специальные категории пользователей, механизмы запрета на отслеживание (do not track) и остальные, не относящиеся явно ни к одной из упомянутых групп [3]. Очевидно, что описания некоторых сценариев могут состоять из похожих слов, и это в некоторых случаях повлечет наложения одних групп на другие.

Исходя из того, что разметка экспериментальных данных включала в себя 10 различных классов, при выделении тем авторы задали число тем, равное 10. Результаты экспериментов, полученных с применением библиотеки Gensim [7] для модели «мешок слов», представлены в табл. 1. В ней показаны полученные кластеры и соответствующие значения координат.

Рассмотрим подробнее полученные результаты. Проанализировав слова-координаты, можно заключить, что темы 1, 2, 5, 7 связаны со сбором куки – фрагментов данных, которые отправляются веб-серверами для оптимизации работы с клиентом и хранятся на компьютере пользователя. Однако можно заметить, что тема 7 связана со сбором куки в рекламных целях (advertis), в то же время семантические модели тем 2 и 5 состоят из одних и тех же слов. Темы 3 и 4 связаны с пере-

Таблица 1

Номер темы	Координата 1	Координата 2	Координата 3	Координата 4
0	0.634"inform"	0.28"may"	0.276"use"	0.232"servic"
1	0.202"cooki"	0.466"inform"	0.336"site"	0.257"use"
2	0.524"privaci"	0.433"polici"	0.388"cooki"	0.219"site"
3	–0.589"servic"	0.344"site"	0.244"parti"	–0.240"third"
4	–0.504"parti"	0.486 "third"	–0.449"servic"	0.235"advertis"
5	–0.594"site"	0.278"cooki"	0.272"websit"	0.264"privaci"
6	–0.326"may"	0.311"site"	0.307"servic"	–0.293"email"
7	–0.437"may"	–0.369"advertis"	0.345"person"	0.319"cooki"
8	0.501"may"	–0.315"email"	–0.281"use"	–0.264"address"
9	–0.488"user"	–0.384"use"	0.310"provid"	–0.301"websit"

дачей информации третьим лицам, причем для темы 4 можно понять, что передача данных третьим лицам осуществляется в рекламных целях (слово advertis). Темы 6 и 8 посвящены сбору контактных данных пользователей, темы 0 и 9 – весьма общие, из их ключевых слов можно утверждать, что речь идет о сборе данных, однако о целях их сбора или их типе сложно судить. Таким образом, можно заключить, что, с одной стороны, очевидно, что в исследуемом наборе данных в основном присутствует информация о сборе персональных данных; с другой стороны, выделить сценарии, отличные от сбора данных, не удалось. При более детальном изучении исходных текстов тестовых данных оказалось, что сделанное предположение верно. Большая часть сегментов политик безопасности посвящена именно сбору данных, а поскольку речь идет об онлайн-соглашениях об использовании данных, то в основном в документах представлена информация о сборе контактных данных и использовании куки. Отдельно следует отметить, что в большинстве случаев сценарий «механизмы запрета на отслеживание (do not track)» связан использованием куки, что опять же подтверждается полученными результатами.

В качестве второй модели векторизованного представления текста была использована модель TF-IDF, которая представляет документ (в данном случае параграф) в форме матрицы, приведенной на рис. 2. Метрика TF-IDF рассчитывается следующим образом:

$$\mathrm{tfidf}(t, d, D) = \frac{n_t}{\sum\limits_k n_k} \times \log \frac{|D|}{\left|\{ d_i \in D : t \in d_i \}\right|}, \quad (1)$$

где D – набор параграфов; n_t – количество употреблений искомого слова в документе; n_k – количество употреблений слова, отличного от искомого. Таким образом, модель TF-IDF придает больший вес словам, которые используются меньшее число раз. Это может быть полезно, когда тексты похожи с точки зрения используемых слов в политике конфиденциальности. Здесь частотные характеристики слов подсчитываются и сопоставляются с параграфами, в которых они встречаются; tfidf – функция подсчета инверсной частотной характеристики, описанная в (1).

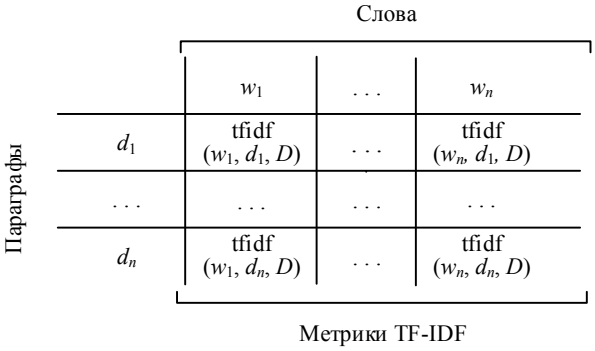

Рис. 2

Результаты экспериментов, полученных с помощью библиотеки Gensim [7] для модели TF-IDF, представлены в табл. 2. В ней, как и в первом эксперименте, представлены десять кластеров и значения атрибутов.

Результаты, с одной стороны, похожи на полученные с помощью модели «мешок слов», в том смысле, что темы снова в большинстве своем посвящены сбору данных первыми лицами. Однако благодаря тому, что TF-IDF придает больший вес словам, которые встречаются реже, наравне с темами, связанными с обработкой куки (темы 0, 1 и 2), появились темы, связанные с обработкой данных детей, подростков (темы 8, 9), темы, касающиеся здоровья (темы 2, 3, 4 и 7) и связанные с изменением содержания политики безопасности (тема 6).

Таблица 2

Номер темы	Координата 1	Координата 2	Координата 3	Координата 4
0	0.202"cooki"	0.2"may"	0.198"inform"	0.198"site"
1	0.573"cooki"	0.262"browser"	0.195"advertis"	0.182"web"
2	–0.406"media"	0.291"cooki"	0.282"health"	0.279"advertis"
3	–0.453"health"	0.258"email"	–0.204"kaleida"	0.191"address"
4	0.423"health"	0.215"media"	0.205"kaleida"	–0.199"secur"
5	–0.299"advertis"	0.262"health"	–0.252"media"	–0.213"privaci"
6	–0.325"media"	0.263"polici"	0.249"privaci"	0.197"chang"
7	0.280"cooki"	–0.216"device"	–0.183"health"	–0.166"social"
8	–0.223"advertis"	–0.206"teenag"	–0.206"inelig"	0.176"child"
9	–0.263" child"	–0.26"wireless"	0.245"message"	0.239"parent"

Таким образом, авторы обнаружили, что формируемые семантические модели не полностью пересекаются с категориями, присутствующими в разметке тестового набора данных. Кроме того, дальнейшие эксперименты показали, что оптимальное число тем, выделяемое алгоритмом латентного семантического анализа, больше, чем количество выделенных категорий. Очевидно, что один и тот же сценарий использования данных описывается сразу несколькими моделями, которые отличаются характеристиками данного сценария. Учитывая все вышеперечисленное, можно заключить, что для эффективного выявления сценариев использования персональных данных для каждого из них необходимо сформировать несколько семантических моделей, причем, возможно, сформированных для различных способов преобразования текста в векторное пространство. Кроме того, следует учесть, что некоторые сценарии использования данных входят в другие сценарии, например сценарий «механизмы запрета на отслеживание (do not track)» – это составная часть сценария «сбор данных первыми лицами», поскольку в нем идет речь о сборе и использовании определенных данных, дающих возможность отследить положение пользователя сервиса, например GPS-координат, куки и т. д.

Применение контекстно-свободных грамматик с тегированием части речи и поиском синонимов. Каждый сценарий использования может быть охарактеризован определенным набором свойств. Например, в сценарии использования персональных данных первыми лицами важно понять, какие данные собираются, с какой целью, каким образом. В сценарии «сроки хранения персональных данных» важно не только какие данные хранятся, но и в какой форме (агрегированные, обезличенные или как есть) и как долго удерживаются. Если латентно-семантический анализ позволяет сформировать модели тем, которые присутствуют в тексте политики безопасности, то применение контекстно-свободного грамматического анализа позволяет выделить свойства сценариев использования персональных данных. В его основе лежит поиск синонимов некоторых ключевых слов и их замена на метки. Например, электронная почта, аватар, местоположение служат объектами и синонимами абстрактной фразы «__CN__», которая означает составное существительное или собираемую тему. Таким образом, все ключевые слова могут быть преобразованы в их значения в контексте некоторой предметной области. Маркировка выполняется элементарно, все слова, которые соответствуют пулу, заменяются меткой этого пула.

Как и в случае с латентно-семантическим анализом, предварительная обработка данных в этом случае включает в себя этап предобработки – токенизацию и лемматизацию для более гибкой замены слов метками из пула синонимов.

Следующим шагом является установление словосочетаний в предложениях, чтобы можно было с уверенностью сказать, что метки пулов синонимов связаны друг с другом и образуют некоторую логическую последовательность. Одним из возможных способов определения словосочетаний в тексте на естественном языке служит синтаксический анализ предложений на основе тегирования по частям речи [8]. После определения частей речи грамматический синтаксический анализатор, входящий в состав библиотеки NLTK [9], основываясь на правилах грамматики, строит дерево предложения. Для предложенного подхода была разработана грамматика

$$\begin{cases} D \to SB \mid SB\,D \mid SB\ U\,D \\ SB \to NPG\ VPG \\ VPG \to VP \mid VP\ VPG \mid VP\ U\ VPG \\ NPG \to NP \mid NP\,NPG \mid NP\ U\,NPG \\ AJPG \to AJ \mid AJ\,AJPG \mid AJ\ U\,AJPG \\ AVPG \to AV \mid AV\,AVPG \mid AV\ U\,AVPG \\ PPG \to PP \mid PP\,PPG \\ VP \to V\,AVPG \mid V\ PPG \mid V\,PPG\,AVPG \\ NP \to NOM \mid DET\,NOM \\ NOM \to N \mid AJPG\,N \\ PP \to NPG \mid P\ NPG, \end{cases} \quad (2)$$

где D – документ; SB – синтаксическая основа предложения с его зависимостями; U – союз; NPG – группа именных фраз; VPG – группа глагольных фраз; AJPG – группа однородных прилагательных; AVPG – группа однородных наречий; PPG – группа однородных дополнений; VP – глагольная группа; NP – именная группа; NOM – номинальная группа; P – предлог; AJ – прилагательное; AV – наречие; PP – существительное с предлогом; N – существительное; V – глагол; DET – определяющее слово.

Грамматика из (2) позволяет авторам рекурсивно восстанавливать грамматические основы предложений и последовательности глаголов, существительных, прилагательных, наречий. Данное решение не идеально, однако оно способно анализировать достаточно сложные предложения.

Поскольку применение данного подхода требует использования пулов синонимов, соответствующих различным ключевым словам, то в грамматику были включены метки пулов синонимов, привязанных к тегам частей речи вручную в нотации NLTK:

$$\begin{cases} U \to NLTK\_CC \\ DET \to NLTK\_DT \\ AJ \to NLTK\_JJ \\ AV \to NLTK\_RB \\ N \to \_\_CN\_\_ \mid \_\_FP\_A\_\_ \mid \_\_TP\_A\_ \mid \\ \qquad NLTK\_N \\ V \to \_\_CV\_\_ \mid NLTK\_V, \end{cases} \quad (3)$$

где NLTK_CC – союзы; NLTK_N – все формы существительных; NLTK_V – все формы глаголов; NLTK_DT – определяющие слова; NLTK_JJ – все формы прилагательных; NLTK_RB – все формы наречий, а теги, начинающиеся с подчеркивания, – это теги пулов синонимов.

Тегирование по частям речи и синтаксический анализ были выполнены с помощью библиотеки обработки естественного языка NLTK [9]. На основе предложенной грамматики, описанной (2) и (3), и маркировки ключевых слов было получено дерево предложения с метками ключевых слов, представленное на рис. 3.

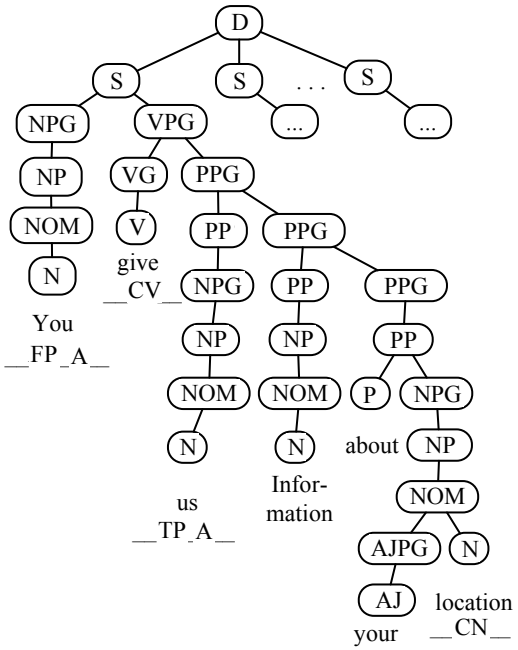

Рис. 3

Здесь __FP_A__ – субъект данных; __TP_A__ – третье лицо; __CV__ – глагол, обозначающий сбор; __CN__ – существительное, обозначающее тип персональных данных. При построении дерева предложений последовательность меток ключевых слов может быть определена. В этом случае на рис. 3 четко видна последовательность __FP_A__, __CV__, __CN__. Такие атомарные последовательности, несущие значения частей предложения, можно агрегировать в список, который описывает смысл всего документа.

Сочетание тегирования ключевых слов и синтаксического анализа позволяет строить деревья, описывающие связи между ключевыми словами. Запросы к таким структурам могут дать информацию, необходимую для построения логической последовательности действующих лиц, их действий, субъектов этих действий и, наконец, различных обстоятельств. Однако очевидно, что для применения данного подхода требуется формирование пула синонимов для каждого свойства сценария, что достаточно трудоемко.

В данной статье представлены результаты исследования двух различных подходов к анализу текстов политик безопасности – латентно-

сематический анализ и подход на основе контекстно-свободных грамматик с тегированием по частям речи. Эксперименты показали, что первый подход может применяться для выявления определенных сценариев использования персональных данных. При этом для каждого сценария необходимо сформировать несколько семантических моделей. Выполненное исследование также показало, что следует переосмыслить типы сценариев использования персональных данных из тестового набора OPP-115 [3], что, возможно потребует определения новых свойств сценариев, и создать собственный размеченный набор данных. В качестве развития этого подхода планируется исследовать различные способы векторизации текстов, а также иные способы построения семантических моделей текстов, в частности, учитывая тесную связь некоторых выявленных сценариев, достаточно перспективным выглядит применение модели коррелированных тем. Подход на основе грамматик с тегированием частей речи позволяет извлечь структурированное, детализированное описание сценария использования персональных данных. Очевидным недостатком представляется необходимость формирования пула синонимов для каждого значимого свойства, данная задача также включена в план дальнейших исследований авторов.

СПИСОК ЛИТЕРАТУРЫ

1. General Data Protection Regulation, GDPR homepage. URL: https://gdpr.eu (дата обращения 14.02.2021).

2. Electronic Passes in Moscow during lockdown. URL: https://www.cnews.ru/news/top/2020-05-25_moskovskij_sajt_s_propuskami (дата обращения 14.02.2021).

3. Pandit H. J., O'Sullivan D., Lewis D. Personalised Privacy Policies. 2018. URL: https://link.springer.com/chapter/10.1007/978-3-030-00063-9_14 (дата обращения 19.02.2021).

4. Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. 2018 / H. Harkous, K. Fawaz, R. Lebret, F. Schaub3, K. G. Shin, K. Aberer. URL: https://arxiv.org/abs/1802.02561v2 (дата обращения 14.02.2021).

5. Novikova E., Doynikova E., Kotenko I. P2Onto: Making Privacy Policies Transparent. Springer, 2020.

6. Landaue T. K., Foltz P. W., Laham D. An introduction to latent semantic analysis // Discourse Proc. 1998. № 25. P. 259–284.

7. Gensim topic modeling library, Gensim homepage. URL: https://radimrehurek.com/gensim (дата обращения 14.02.2021).

8. Automated Extraction of Vulnerability Information for Home Computer Security / S. Weerawardhana, S. Mukherjee, I. Ray, A. Howe. Springer, 2015. P. 356–366. URL: https://link.springer.com/chapter/10.1007/978-3-319-17040-4_4 (дата обращения 14.02.2021).

9. Nateral Language ToolKit, Analyzing Sentence Structure. NLTK homepage. URL: https://www.nltk.org/book/ch08.html (дата обращения 14.02.2021).

M. D. Kuznetzov, V. S. Myadzel
Saint Petersburg Electrotechnical University

E. S. Novikova
St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS)

TOWARDS APPLICATION OF TEXT MINING TECHNIQUES TO THE ANALYSIS OF THE WEBSITE'S ONLINE CONSENTS

Website's online privacy consents provide end users information about how they personal data collected, processed and shared with third parties by web services. However, in major cases they are written in unclear and not transparent manner. In the paper, the authors investigate application of two different text mining approaches to analyze texts of privacy policies. To detect different personal data usage scenarios, latent semantic analysis technique with several statistic text models was applied. Also, to establish links between elements of data scenario, the part-of-speech based analysis was used. Using the part-of-speech based analysis, it is possible to reveal logical sequences which describe data practices. The authors tested the selected algorithms against a set of labelled privacy policies, generated within Usable Privacy Project, and discuss the obtained results.

Text analysis, latent semantic analysis, synonym search, context-free grammars, natural language processing, internal data convention, morphological analysis